Information Fusion 108 (2024) 102372

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Transformer-based self-supervised image super-resolution method for Rotating Synthetic Aperture system via multi-temporal fusion

Yu Sun [a], Xiyang Zhi [a], Shikai Jiang [a,*], Guanghua Fan [b,*], Tianjun Shi [a], Xu Yan [c,*]

[a] Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China
[b] Department of Optoelectronics Science, Harbin Institute of Technology at Weihai, Weihai, Shandong 264209, China
[c] College of Science and Engineering, University of Glasgow, Glasgow, G12 8QQ, UK

ARTICLE INFO

Dataset link: HRRSD

Keywords:
Image super-resolution
Vision Transformer
Optical remote sensing
Self-supervised
Rotating synthetic aperture
Wavelet fusion

ABSTRACT

Rotating Synthetic Aperture (RSA) technology is one of the distinctly advantageous Earth geostationary orbit optical remote sensing technologies. However, the continuous rotation of the RSA system's rectangular primary mirror results in a discernible drop in resolution along the shorter side of the mirror. Additionally, the captured images exhibit periodic and time-varying characteristics. To improve the image quality to meet interpretation needs, we first delineate the imaging process of the rotating primary mirror and analyze the characteristics of image degradation based on the system's imaging mechanism. Then, we propose a dual super-resolution (SR) framework based on Swin Transformer and introduce a self-supervised learning method for jointly training the unified SR network using wavelet fusion. The self-supervised learning method effectively utilizes the spatiotemporal correlation of the information contained in images captured at different rotation directions of the rectangular pupil. Moreover, the attention mechanism in Transformer can adopt a global perspective and utilize content-based interactions between image content and attention weights to model strong long-range dependencies in remote sensing images. This approach significantly enhances image quality along the pupil's shorter side, consequently yielding superior results. Extensive digital and semi-physical imaging experiments, involving six aspect ratios of the primary mirror, demonstrate that our SR method surpasses state-of-the-art methods. The work in this paper can serve as a valuable reference for future space applications of the RSA technology.

1. Introduction

Optical remote sensing satellites have significant advantages, such as wide area coverage, fixed-point monitoring, and high temporal resolution. Obtaining high spatial resolution images depends primarily on a large-aperture primary mirror. In geostationary orbit, optical apertures larger than 10 m are theoretically necessary to achieve optical imaging with a spatial resolution of 1–2 m [1–3]. Traditional optical systems with large aperture single mirrors are limited by weight, volume, complexity, and the rocket's carrying capacity, making it difficult to meet the requirements of low-cost applications and lightweight designs. To achieve a large aperture while meeting the rocket's carrying capacity, several new imaging technologies have emerged. These include sparse aperture imaging technology, membrane diffraction imaging technology, and rotating synthetic aperture (RSA) imaging technology. The technique of sparse aperture imaging can reduce the aperture by splicing small-aperture mirrors into one larger aperture primary mirror to some extent. However, it requires a precise on-orbit deployment structure and real-time detection and correction of the overall surface error, resulting in a significant increase in technical complexity [4–7]. The membrane diffraction imaging technique has the benefits of being lightweight and foldable, but due to the coupling effect of thin film material and diffraction mechanism, the imaging quality is unpredictable and often unable to meet the high-quality optical imaging standards [8–10]. The RSA technique, as shown in Fig. 1, employs a rectangular primary mirror. The mirror achieves a resolution and image quality comparable to that of a system with a circular aperture of equivalent caliber during rotation. The system offers numerous advantages, including a lightweight primary mirror and eliminates the need for splicing or maintaining surface shape. These advantages significantly reduce the difficulty of overall system design, manufacturing, and detection [11,12]. The RSA technique stands out as a leading system design in contemporary high-resolution space optical imaging, emerging as a pivotal development direction for future geostationary orbit high-resolution optical cameras. Nevertheless, during the imaging

* Corresponding authors.
E-mail addresses: jiangshikai@hit.edu.cn (S. Jiang), fangh@hitwh.edu.cn (G. Fan), 2703613Y@student.gla.ac.uk (X. Yan).

https://doi.org/10.1016/j.inffus.2024.102372
Received 4 February 2024; Received in revised form 10 March 2024; Accepted 19 March 2024
Available online 24 March 2024

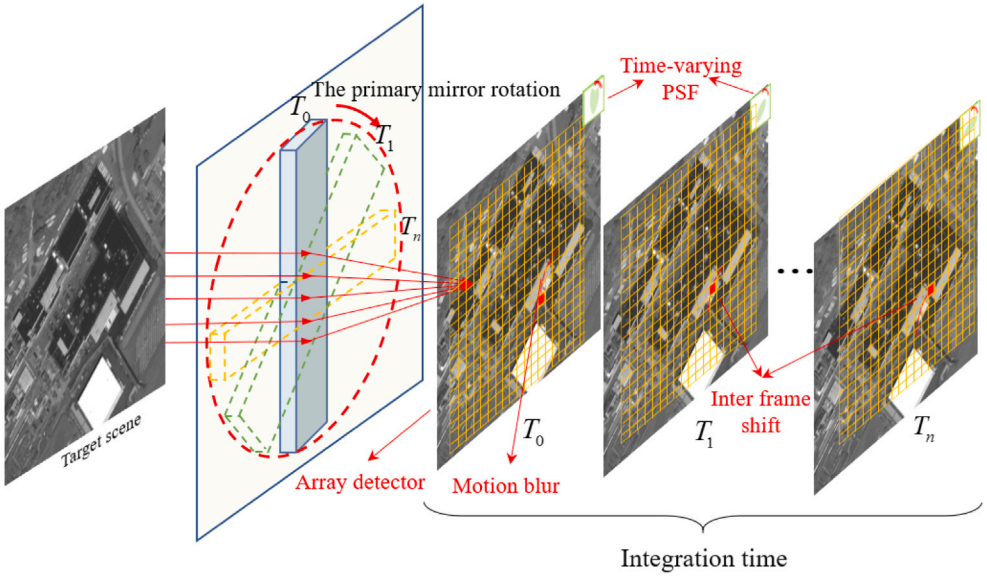

Fig. 1. Imaging principle of RSA. In contrast to traditional imaging methods, the key distinction lies in its use of a continuously rotating rectangular mirror as the primary mirror.

process, the continuous rotation of the rectangular primary mirror causes unique degradation characteristics in image quality. Thus, it is necessary to combine image processing methods to enhance the imaging quality [13,14].

The modulation transfer function compensation technique has always been an important component of practical applications in optical remote sensing satellite systems. In earlier years, a series of high-resolution remote sensing satellites, such as IKONOS, QuickBird, GeoEye, and WorldView, have all adopted this method to enhance the quality of image products [15,16]. Deep learning has the ability to adaptively learn the mapping relationship between high-resolution and low-resolution information. This feature renders image super-resolution (SR) methods based on deep learning significantly superior to traditional approaches, thereby establishing them as the prevailing methods in recent years. It can provide a reference for remote sensing image SR. The single-image super-resolution (SISR) methods based on deep learning can be roughly divided into two categories [17]: the first category is explicit methods based on classic degradation models or their variants. Examples include SRGAN [18], ELAN [19], Omni-SR [20], HAT [21], and Real-ESRGAN [22] that utilize external datasets, as well as KernelGAN [23], DualSR [24], and DBPI [25] that exploit internal statistics of patch recurrence. The second category is implicit methods that learn the data distribution within external datasets, such as CinCGAN [26] and FSSR [27]. However, the RSA system possesses unique imaging characteristics, and its image degradation mechanism is more complex compared to traditional systems. Therefore, directly applying generic SISR methods would encounter the following challenges: (1) During the imaging process, the anisotropic point spread function (PSF) of the system continuously changes with rotation, resulting in images exhibiting temporal periodicity and spatial asymmetry. Existing SR methods do not consider this unique imaging characteristic, which may lead to suboptimal performance. (2) Deep learning is generally known as a data-driven approach. In other words, the effectiveness of deep learning methodologies is strongly tied to both the quantity and quality of the available data. However, constructing a suitable large external dataset for the RSA system, which has not yet been implemented for in-orbit applications, is relatively challenging. (3) The above SR methods are mostly based on convolutional neural networks (CNN), but strong long-range dependencies of remote sensing images makes the effectiveness of CNNs with local inductive bias difficult to meet application requirements. Therefore, we aim to investigate image SR methods tailored specifically for the RSA system, based on its degradation mechanism and imaging characteristics. Building upon this foundation, our objectives encompass two main aspects: firstly, employing self-supervised learning techniques to utilize the high-resolution information inherent in images to guide the reconstruction of the low-resolution direction of the images without the need for large external training datasets; secondly, exploring model architectures based on Vision Transformer to enhance the effective utilization of internal long-range dependencies within remote sensing images. Moreover, for innovative imaging systems like the RSA, additional physical or semi-physical imaging experiments are required to further validate the performance of the SR methods.

In order to address the issue, we first conduct an analysis of the asymmetric spatial distribution characteristic and time-varying characteristic of the PSF, in accordance with the imaging mechanism of the system. Subsequently, we propose a SISR method based on Swin Transformer [28]. Finally, digital simulation and semi-physical imaging experiments demonstrate that our proposed method can enhance image resolution while correcting spatial anisotropy and improving texture detail clarity. Under conditions where the aspect ratio of the rectangular primary mirror ranges from 3 to 7, the performance of our method surpasses that of state-of-the-art generic SISR methods across various target scenes such as residential areas, harbors, forests, storage yards, and airports.

The contributions of this study can be succinctly outlined as follows:

1. Analysis of the degradation characteristics of image quality in the RSA imaging system, specifically focusing on temporal periodicity and spatial anisotropy.
2. Proposal of a self-supervised learning method based on wavelet fusion that corresponds to the phenomenon of image resolution being significantly higher in the long edge direction of the rectangular primary mirror than in the short edge direction. This eliminates the need for large external datasets in model training.
3. Introduction of a dual SR framework to utilize content-based interactions between attention weights and image content unique to the Swin Transformer. This enhances the ability to exploit internal long-range dependencies within remote sensing images, compensating for the significant information loss along the shorter side direction.

4. Verification of our approach through digital simulation and semi-physical imaging experimentation to illustrate its superiority.

The paper is organized as follows: Section 2 provides an overview of the current research status. In Section 3, the imaging mechanism of the system is examined, with an analysis of its characteristics leading to image quality degradation. Subsequently, we introduce a self-supervised SR method tailored for the RSA system. The efficacy of our approach is substantiated through digital simulation and semi-physical simulation experiments in Section 4. Finally, our conclusions are summarized in Section 5.

2. Background

The task of SISR involves the recovery of a high-resolution image I_H from its corresponding low-resolution counterpart I_L. The relationship between these two variables is defined by the classical degradation model, which can be expressed as:

$$I_L = \left(I_H * k\right)_{\downarrow} + n, \tag{1}$$

where k represents the blur kernel, $*$, n, and $\downarrow$ denote the convolution operator, additive noise, and the down-sampling operator, respectively.

Existing SISR methods can be broadly categorized into two groups: explicit methods, which rely on classic degradation models or their variations, and implicit methods, which utilize data distribution from external datasets. The fundamental concept of explicit methods is to learn the blur kernel k and additive noise n in the classical degradation model directly from the training data. The representative approaches include SRGAN [18], Omni-SR [20], HAT [21], and Real-ESRGAN [22], which exploit external datasets. Another set of approaches suggests exploiting the internal statistics of patch recurrence, like DualSR [24], DBPI [25], and KernelGAN [23]. However, the RSA system's unique timeseries imaging mode makes it difficult for a single deep neural network to accurately estimate the asymmetric kernel with time-varying characteristic, that is, the method based on estimating explicit parameters such as blur kernel is no longer applicable to the system. Implicit methods, on the other hand, do not rely on explicit parameterization. Instead, they typically learn the underlying SR model implicitly from the data distribution within external datasets. The representative approaches include CinCGAN [26] and FSSR [27]. However, implicit methods rely on external datasets and may not produce satisfactory results for images with degradations beyond their training sets [29,30]. This is particularly challenging for new imaging systems like the RSA system, which has not been deployed in orbit yet, making it harder to create a sufficiently large and relevant external dataset. In addition, external information is subject to a higher degree of uncertainty compared to internal information. As a result, methods that rely solely on external datasets for full supervision more commonly generate significant artifacts in the SR results compared to self-supervised methods based on the specific internal information of the image itself. These artifacts pose a significant disadvantage to the space application of the RSA imaging technology.

The aforementioned SR methods are mostly based on CNNs. However, recently, Transformer has emerged as a new possibility for computer vision, thanks to its exceptional ability to model long-range dependencies. Remote sensing images, in contrast to natural images, tend to be larger in size and have more objects with varying sizes and orientations. Moreover, remote sensing images have a higher information density. Vision Transformer can achieve global dependencies by facilitating interactions among arbitrary pixels in remote sensing images. Furthermore, the hierarchical Vision Transformer, which combines the hierarchical idea in CNNs, can maintain modeling power while reducing the computational cost to complexity linearly proportional to the image size [31,32]. Consequently, the general paradigm of network design has gradually shifted from CNNs to Transformer. Transformer-based visual models have not only achieved results comparable to or even surpassing CNNs in tasks such as image super-resolution, image restoration, semantic segmentation, point cloud analysis, object detection, and recognition in natural images [33–44], but their applications in remote sensing images are also becoming increasingly widespread. [45] conducted an empirical study on remote sensing pretraining and obtained Transformer-based pretrained backbones that have demonstrated promising performance in processing remote sensing images. The study also investigated the impact of these pretrained backbones on representative downstream tasks. Subsequently, [46] introduced a large vision model customized for remote sensing tasks, incorporating a rotated varied-size window-based attention mechanism. This model excels not only in computational complexity and data efficiency during transfer but also demonstrates competitive performance in downstream tasks. For research on remote sensing image SR, [47] utilizes the self-attention mechanism to design a module for single-scale self-similarity exploitation, enabling the computation of feature correlations within the same scale. Additionally, to capture repetitive structures across different scales, it incorporates a cross-scale connection structure. The combination of these two components allows for the simultaneous utilization of both single-scale and cross-scale similarities. [48] enhances the SR performance of remote sensing images further by employing a Transformer-based multistage enhancement structure. This structure integrates multi-scale high-dimensional and low-dimensional features and captures long-range dependencies between them.

3. Methodology

3.1. Overview

The RSA system, as shown in Fig. 1, is an imaging system that uses a large aspect ratio primary mirror to rotate during operation. Firstly, we conduct a thorough analysis of the spatial asymmetry and temporal periodicity characteristic of the PSF specific to the system. In light of this unique imaging mechanism, we present a self-supervised learning method that employs wavelet fusion. Additionally, we introduce a dual SR network based on Transformer. The network leverages the content-based interactions between attention weights and image content exclusive to Swin Transformer, thereby improving the ability to utilize the internal information of remote sensing images. Further, the attention mask module is removed to enhance the residual Swin Transformer block's ability to model stronger long-range dependencies of remote sensing images. Fig. 2 illustrates the overall process.

3.2. Mechanism analysis of the RSA system's imaging characteristics

The imaging quality of the acquired image at any time in the direction of the long side of the primary mirror is comparable to that of the single circular primary mirror with the same caliber. As the primary mirror rotates, the high-quality imaging produced by the long side covers all directions over time, effectively achieving the equivalent of a larger-caliber single primary mirror through time division.

The rectangular pupil function is shown in Fig. 3. The PSF at the same time (as shown in Fig. 4) can be obtained by taking the Fourier transform of the rectangular pupil function and then taking the square of the modulus:

$$\begin{aligned} PSF_{\text{rect}}(x,y,t) = ab \cdot \operatorname{sinc}(a(x\cos(\omega t+\varphi_0) - y\sin(\omega t+\varphi_0))) \\ \times \operatorname{sinc}(b(x\sin(\omega t+\varphi_0) + y\cos(\omega t+\varphi_0))), \end{aligned} \tag{2}$$

where a and b are, respectively, the length and width of the rectangular pupil, ω is the rectangular primary mirror rotation angular velocity, and $\omega t + \varphi_0$ is the rotation phase.

Referring to Eq. (2), the PSF at time t is determined by the rotation angle and aspect ratio of its rectangular pupil, exhibiting the following characteristics:

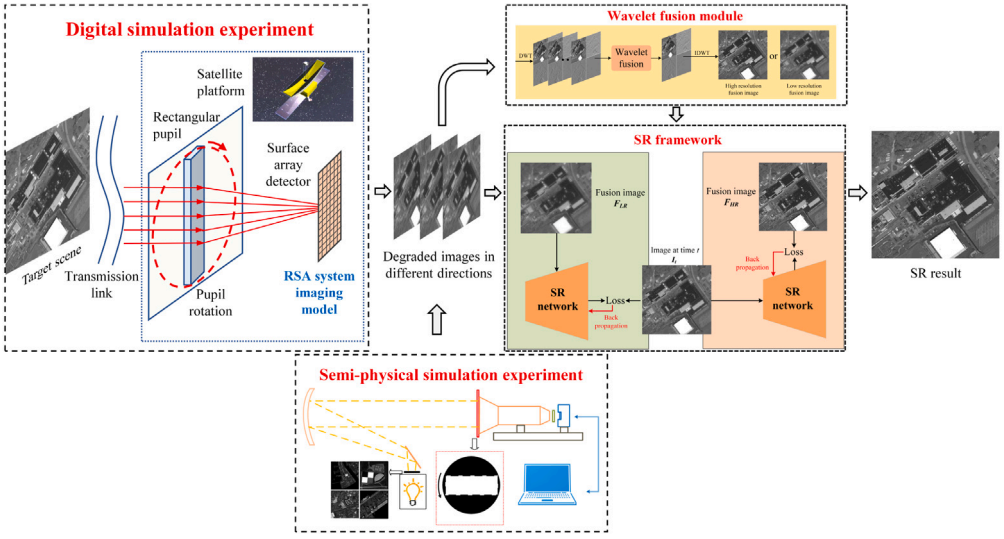

Fig. 2. Overview Flow Chart: The image super-resolution reconstruction part consists of a dual SR frame, which comprises a wavelet fusion module and a Transformer-based network. Validation experiments include digital simulation and semi-physical imaging experiments.

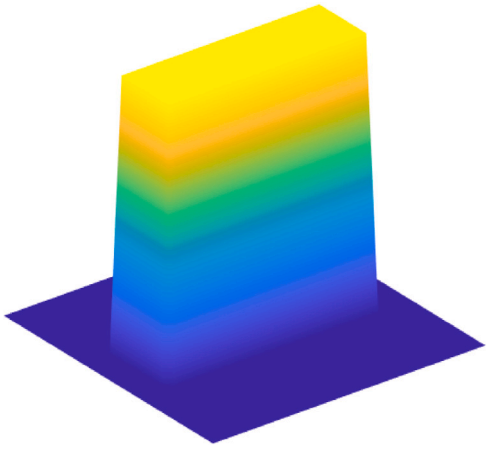

Fig. 3. 3D schematic of the rectangular pupil function.

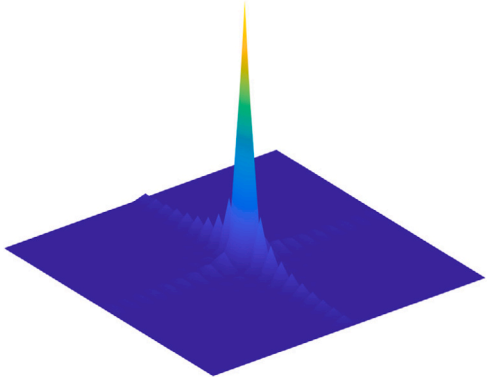

Fig. 4. 3D schematic of the point spread function (PSF).

1. The PSF of a rectangular aperture approximates an elliptical shape when the secondary diffraction effects are disregarded. This elliptical form is determined by the aspect ratio of the rectangular aperture.
2. The PSF of the rectangular pupil undergoes temporal variations, with the short axis aligning in the direction of the long side of the pupil.

For instance, when the aspect ratio is 3, the PSFs at various rotation angles are shown in Fig. 5, and the corresponding degraded images are presented in Fig. 6. It is apparent that there is a significant variation in image resolution across different directions, with notably lower resolution observed on the shorter side compared to the longer one.

Specifically, due to the continuous rotation of the rectangular mirror during the imaging process, the PSF changes over time. The PSF is different at each time, which leads to complex nonuniform degradation of the acquired images, and the resolution is influenced by both the aspect ratio and the rotation speed of the rectangular mirror. Finally, due to the influence of detector sampling, the resolution decreases further. Therefore, it is imperative to develop an SR model tailored to enhance the resolution of images obtained through the RSA system, aligning with the system's imaging characteristics. This model should leverage the spatial correlation inherent in the image to enhance resolution in its low-resolution direction.

3.3. Self-supervised learning method based on wavelet fusion

In most studies on learning-based image SR, the paired training dataset is created by downscaling high-resolution images with a predetermined operation (e.g., bicubic) [49]. However, estimating the blur kernel of the RSA imaging system is too complex and kernel mismatch may produce unwanted artifacts. To address this issue, the self-supervised method utilizes the internal statistics of an image as a useful prior for solving unconstrained problems like image SR [50]. Wavelet transform is a space-frequency domain analysis method used to decompose an image into a combination of average and detail images, each representing distinct structures of the image. This facilitates the extraction of both the overall structure and intricate details, enabling the construction of high-resolution-low-resolution (HR-LR) image pairs for

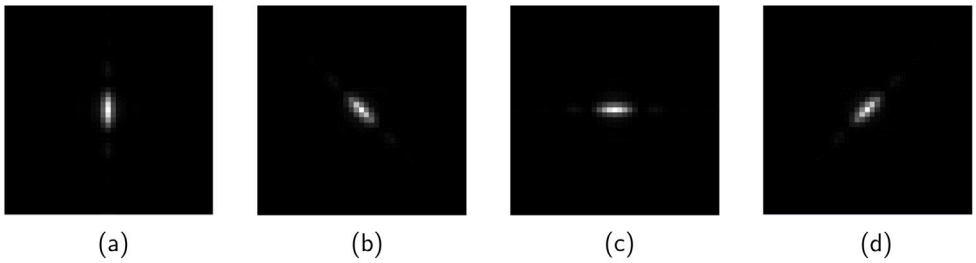

Fig. 5. PSFs at various rotation angles. (a) 0°, (b) 45°, (c) 90°, (d) 135°.

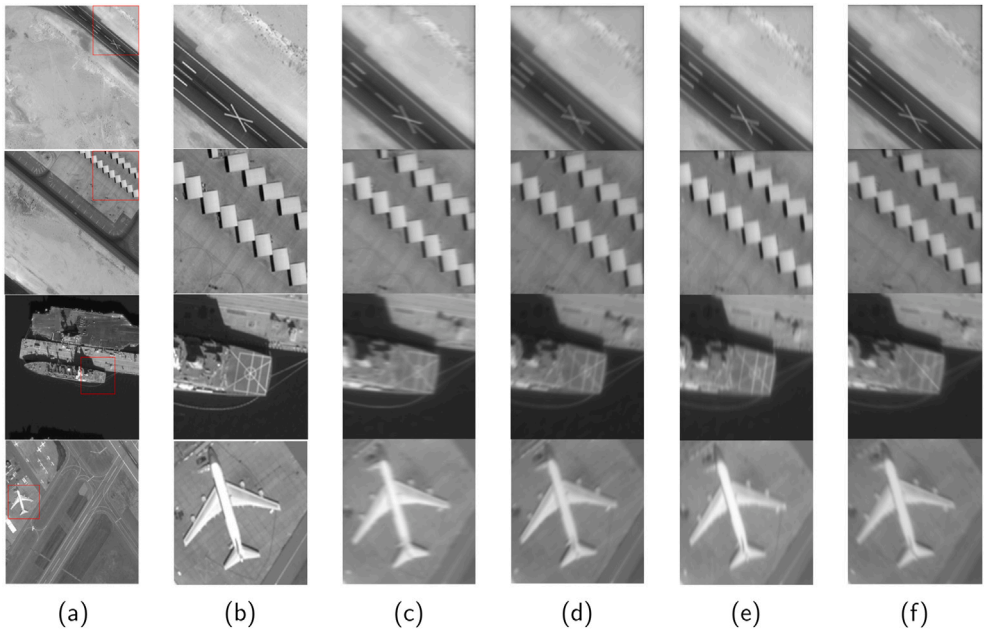

Fig. 6. Image degradation with varying rotation angles. (a)–(b) Example image scene. (c)–(f) Degraded images at various rotation angles.

self-supervised training. The formula of the two-dimensional discrete wavelet transform is:

$$DWT(j,m,n) = 2^j \sum_x \sum_y I(x,y)\psi(2^j x - m, 2^j y - n), \quad x,y \in Z, \tag{3}$$

where I represents the input image to be decomposed, j is the scale of wavelet decomposition, ψ is the wavelet basis function, and m and n represent the translation parameters that determine the position of the wavelet function ψ in relation to the image position $I(x,y)$.

Fig. 7 exhibits the wavelet decomposition outcomes of the resolution target image, acquired by semi-physical imaging experiments. The outcomes are categorized into four sections: Region A corresponds to the subsampling of the primary image, representing the low-frequency part. Regions H, V, and D represent the details (high-frequency) of the original image's rows, columns, and diagonal directions, respectively. It is apparent that the resolution in the long side direction varies from that in the short side direction.

Drawing on the aforementioned analysis, as shown in Fig. 8, we adopt a self-supervised learning approach relying on wavelet fusion, which follows the subsequent steps:

1. Decompose the sequence images $I_1, I_2, \ldots, I_n$ which acquired in one rotation cycle of the pupil.
2. Take the weighted average of the wavelet coefficients of the low-frequency part of the sequence images $I_1, I_2, \ldots, I_n$ as the

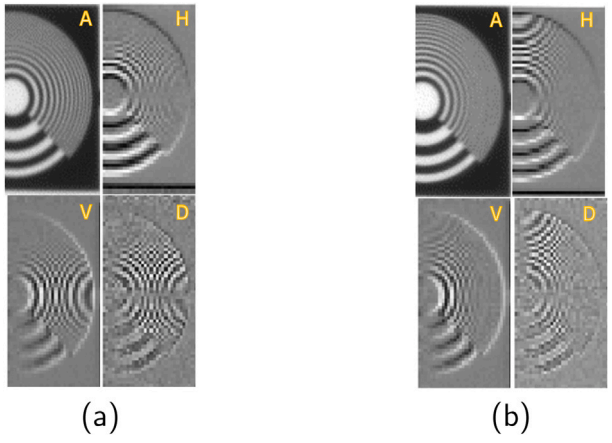

Fig. 7. Wavelet decomposition results of the resolution target image obtained by semi-physical imaging experiments. (a) The results with the rotation angle of 0°. (b) The results with the rotation angle of 90°.

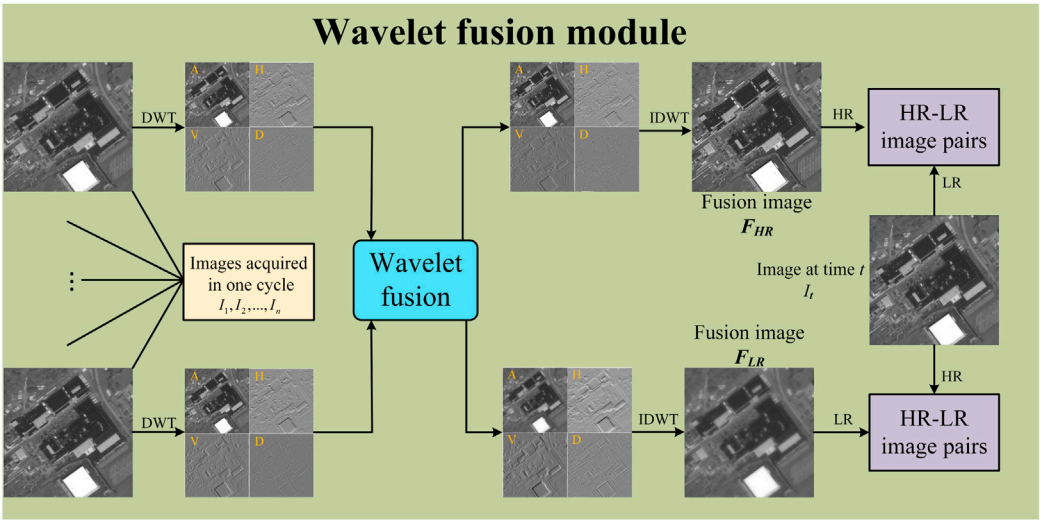

Fig. 8. The architecture of the self-supervised learning method based on wavelet fusion. The wavelet fusion module can fuse the information of high/low resolution in different directions of the images obtained during one rotation cycle to obtain fusion images, which are then used to construct HR-LR image pairs.

wavelet coefficients of the low-frequency part of the fusion image.

3. Compare the wavelet coefficients of the three high-frequency components of the sequence images $I_1, I_2, \ldots, I_n$, then take the wavelet coefficients with the largest (or smallest) absolute value as the wavelet coefficients of the fusion image F_{HR} (or F_{LR}).
4. After determining the wavelet coefficient, take the inverse wavelet transform to obtain the fusion image, then upsample (or downsample) the fusion image F_{HR} (or F_{LR}).

We choose the Haar wavelet and use the 2-D fast wavelet transform [51]. By taking the wavelet coefficient with the largest/smallest absolute value of the high-frequency parts, the information of high/low resolution in different directions can be respectively fused by the wavelet to obtain fusion images. These fusion images are then combined with the original images captured by the RSA system to construct HR-LR image pairs, which will be used to train the SR network.

3.4. Dual self-supervised learning-based image SR framework

The self-supervised method, relying on the internal statistics of remote sensing images themselves, obviates the need for an extensive collection of external training datasets, and it is not prone to producing "hallucinations" or "artifacts" out of nothing. Therefore, we propose a unified self-supervised remote sensing image SR framework that comprises a wavelet fusion module for constructing HR-LR image pairs and an SR network. As shown in Fig. 9, on one side of the framework, the wavelet fusion module is utilized to combine the low-resolution information of the images obtained within one rotation cycle of the pupil, producing the LR image in the HR-LR pair. This LR image is then restored by the SR network, with the original single-frame image acquired by the RSA system acting as HR for supervision. In other words, on this side, the loss functions used for backpropagation include pixel loss and edge loss, calculated between the super-resolved image obtained through the SR network's processing of the input LR image (F_{LR}) and a single-frame image acquired by the system (I_t). The pixel loss L_1 uses L1 distance, defined as:

$$L_1\left(F_{LR}, I_t\right) = \left|S\left(F_{LR}\right) - I_t\right|_1, \tag{4}$$

where $S(\cdot)$ represents the SR network.

The edge loss L_{edge} is defined as:

$$L_{\mathrm{edge}}\left(F_{LR}, I_t\right) = \sqrt{\left\|\Delta\left(S\left(F_{LR}\right)\right) - \Delta(I_t)\right\|_2^2 + \varepsilon^2}, \tag{5}$$

where Δ represents the Laplacian operator and the constant ε is empirically set to 0.001.

The total loss function L is as follows:

$$L\left(F_{LR}, I_t\right) = \lambda_1 \cdot L_1 + \lambda_2 \cdot L_{\mathrm{edge}}, \tag{6}$$

where λ_1 and λ_2 are set as 1 and 0.05 empirically.

In parallel, on the opposite side of the framework, the wavelet fusion module is employed to merge the high-resolution information from the images acquired within one rotation period, producing the HR image in the HR-LR pair. Subsequently, the SR network is utilized to restore the original single-frame image obtained by the system. Accordingly, the loss is as follows:

$$L\left(I_t, F_{HR}\right) = \lambda_1 \cdot \left|S\left(I_t\right) - F_{HR}\right|_1 + \lambda_2 \cdot \sqrt{\left\|\Delta\left(S\left(I_t\right)\right) - \Delta(F_{HR})\right\|_2^2 + \varepsilon^2}. \tag{7}$$

It is noteworthy that the HR images on both sides of the framework have a higher resolution than the LR images in certain directions. This indicates that the SR network on both sides can be trained jointly using the HR-LR pairs.

The SR network, denoted as S, is constructed based on SwinIR [52], as illustrated in Fig. 10. The network utilizes only a small number of convolutional layers in both the shallow feature extraction and high-quality image reconstruction modules. Specifically, for a low-resolution input image LR, the shallow feature extraction process employs a 3×3 convolutional layer, denoted as $E_{SF}(\cdot)$, to extract the shallow feature F_{SF} as follows:

$$F_{SF} = E_{SF}(LR). \tag{8}$$

Next, the deep feature extraction module $E_{DF}(\cdot)$ is employed to extract the deep feature F_{DF} as follows:

$$F_{DF} = E_{DF}\left(F_{SF}\right). \tag{9}$$

Finally, the reconstruction module $R_{HQ}(\cdot)$ utilizes a single convolution layer to aggregate shallow and deep features. It also employs a sub-pixel convolution layer to upsample the features. The reconstruction

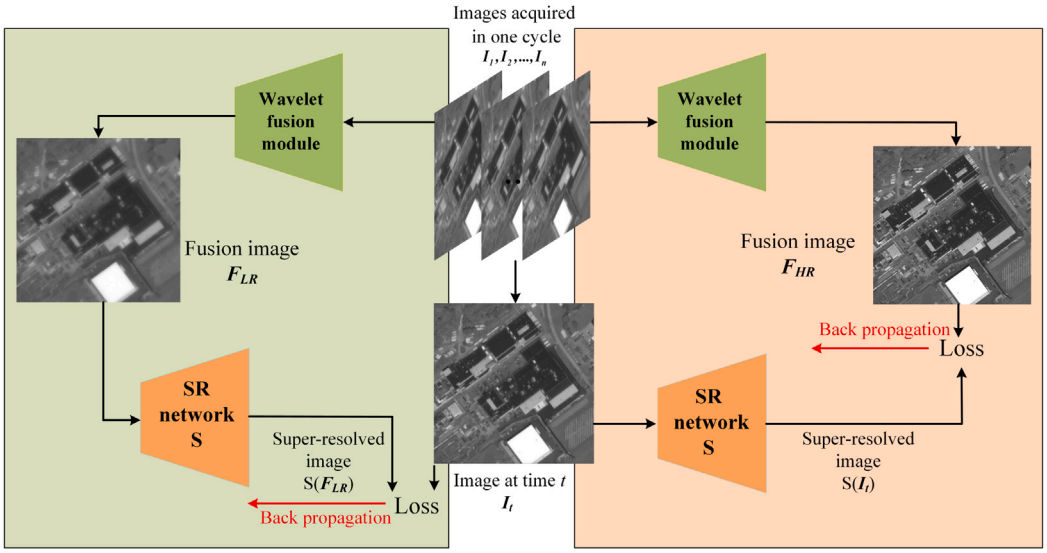

Fig. 9. Overall framework of the proposed dual SR Framework. Each side of the framework includes a wavelet fusion module and a Transformer-based SR network. The SR networks on both sides can be trained jointly.

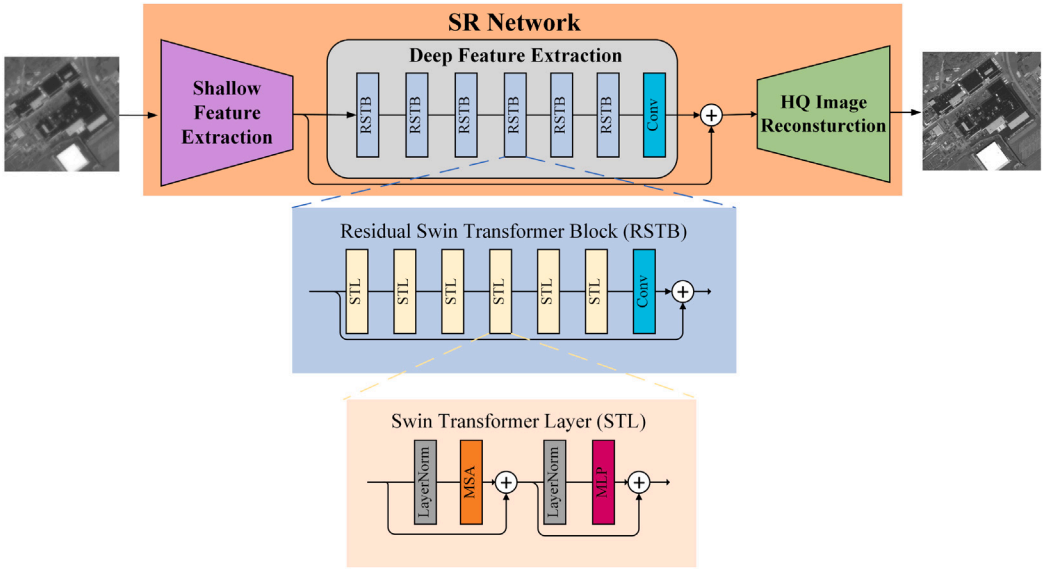

Fig. 10. The SR network structure.

process is performed as follows:

$$S(LR) = R_{HQ}\left(F_{SF} + F_{DF}\right). \tag{10}$$

CNNs do not process information about every pixel in the input image but instead perceive local regions due to their local inductive bias. This property of convolution makes it ineffective for establishing long-range dependencies in space, so convolutional layers occupy a minimal proportion in our SR network. On the contrary, the Transformer-based deep feature extraction module utilizes content-based interactions between attention weights and image content, differing from convolution and enabling better modeling of long-range dependencies. More specifically, in the deep feature extraction module, only the final layer is a 3×3 convolutional layer. The preceding layers consist of P Swin Transformer blocks (RSTB). These blocks are used to extract intermediate features $F_1, F_2, \ldots, F_P$, block by block, as follows:

$$\begin{aligned} F_1 &= RSTB_1\left(F_{SF}\right), \\ F_i &= RSTB_i\left(F_{i-1}\right), \\ F_{DF} &= Conv\left(F_P\right), \end{aligned} \tag{11}$$

where $i = 2, 3, \ldots, P$ and $Conv$ denotes the last convolutional layer.

RSTB comprises Swin Transformer layers (STL) [28] integrated with convolutional layers. The shifted window mechanism of STL is conducive to modeling long-range dependencies, aiding in the reconstruction in the lower resolution direction by utilizing high-resolution

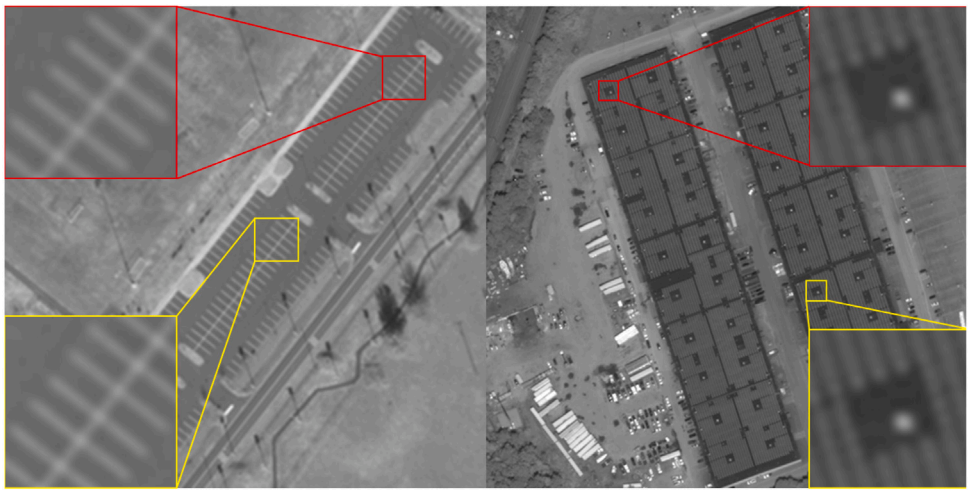

Fig. 11. Long-range dependencies: The recurrence frequency of a patch remains high even at greater spatial distances in remote sensing images.

information retained from other directions in the image itself. Additionally, for remote sensing images, as illustrated in Fig. 11, the recurrence frequency of a patch remains high even at a greater spatial distance. Therefore, to effectively leverage this characteristic of remote sensing images, we remove the attention mask during the computation of self-attention in the shifted window-based attention module of STL, capturing the stronger long-range dependencies present in remote sensing images.

The proposed SR framework is inspired by CycleGAN [53] and DBPI [25], but there are two significant differences. Firstly, the core of the SR framework is the self-supervised Swin Transformer, not the generator and discriminator in a generative adversarial network. One reason for this is that remote sensing images have a lot of internal information redundancy (e.g., recurrence of small patches), which can produce useful specific image priors [50]. Additionally, image patches exhibit a higher frequency of occurrence within a single remote sensing image, thereby enhancing the potency of internal image-specific statistics compared to generic external statistics. Unlike the local inductive bias in convolution, Transformer can take a global perspective and exploit correlations between pixels, making it possible to effectively utilize high-resolution information in different directions of the target scene. Secondly, the training set is obtained through wavelet fusion, rather than being generated by a deep neural network. Using a network to generate training data presents two challenges: the blur kernel is time-varying, and the network may be trained to generate easily downscale and recoverable images, such as linear interpolation. By creating the paired training dataset through wavelet fusion, a trivial solution is avoided.

3.5. Computational complexity analysis

For an RGB image of size $h\times w\times 3$, wavelet transform first requires $\log_2 h$ and $\log_2 w$ iterations for each row and column of the image. Each iteration divides the row/column into sub-rows/columns with half the original length, and performs wavelet transform on each sub-row/column. The subsequent fusion process requires iterating through all the elements of the coefficient matrix one by one. Finally, considering the input channel is 3, the total computation complexity of wavelet fusion is:

$$\Omega_{\mathrm{WF}} = 3hw(\log_2 h + \log_2 w + 1). \tag{12}$$

Swin Transformer performs self-attention computation within a set of local windows. Initially, the input features $\mathrm{X} \in \mathbb{R}^{h\times w\times C}$ are divided into non-overlapping windows, with each window containing $M\times M$ patches. The computational complexity of window-based self-attention is:

$$\Omega_{\mathrm{WMSA}} = 4hwC^2 + 2M^2hwC. \tag{13}$$

4. Experiments

4.1. Experimental configuration

To assess the efficacy of the proposed SR method, we carry out digital simulation as well as semi-physical imaging experiments. The digital simulation experiment is conducted with high-resolution remote sensing images using a full-link digitization mode to simulate the system's image quality degradation process [54,55]. According to the analysis in Section 3, taking the target scene image $I_0(x, y)$ as input, the degraded image can be represented as follows:

$$\begin{aligned} I(x, y, t) &= \frac{1}{T}\int_0^t I_0(x, y) * PSF_{\mathrm{link}}\,\mathrm{d}t, \\ PSF_{\mathrm{link}} &= PSF_{\mathrm{ele}} * PSF_{\mathrm{det}} * PSF_{\mathrm{opt}} * PSF_{\mathrm{rect}} * PSF_{\mathrm{atm}}, \end{aligned} \tag{14}$$

where PSF_{rect} is the PSF of rectangular pupil in Eq. (2), PSF_{ele}, PSF_{det}, PSF_{opt}, and PSF_{atm} represent the PSF of electronic system, detector, optical system defocusing, and atmospheric disturbance respectively. Table 1 displays the pertinent simulation parameters.

For the semi-physical imaging experiment, an imaging experiment platform is constructed to simulate the RSA imaging process, as illustrated in Fig. 12. A rotating rectangular pupil optical element is added to the front of a high-quality optical lens to mimic the dynamic imaging process. A spectral filter is included to simulate the influence of varying spectral widths. The data acquisition, processing, and analysis of images are carried out by a computer. High-resolution images are used in the experiments, and these images are produced by sophisticated cartographic equipment [56]. Fig. 13 presents some of the experimental components.

We conducted a quantitative evaluation of the proposed SR method using the HRRSD dataset [57]. HRRSD comprises a total of 21,761 images acquired from Google Earth and Baidu Map. Additionally, we also utilized images from the WorldView-3 (WV3) satellite. These images contain target scenes with varying texture richness, including airports, ports, residential areas, forests, and farmlands, as shown in Fig. 14. This additional dataset was employed for further evaluation of the proposed method.

Table 1
Simulation parameters.

Operation	Name	Value
Scene and atmospheric transmission	Scene	High-resolution radiance images
	Atmospheric path radiation	Standard atmospheric model
	Atmospheric transmittance	Standard atmospheric model
	Aerosol	Standard atmospheric model
	Scattering and absorption	Standard atmospheric model
Camera	Aspect ratio of the rectangle primary mirror	3~8
	Focal length	50 m
	Equivalent diameter of the primary mirror	6 m
	Center wavelength	500 nm
	Rotational angular velocity	0.01~0.05 rad/s
Signal transmission and conversion	Pixel size	9 μm
	Depletion width	5 μm
	Diffusion length	2 μm
	Integration time	0.01 s

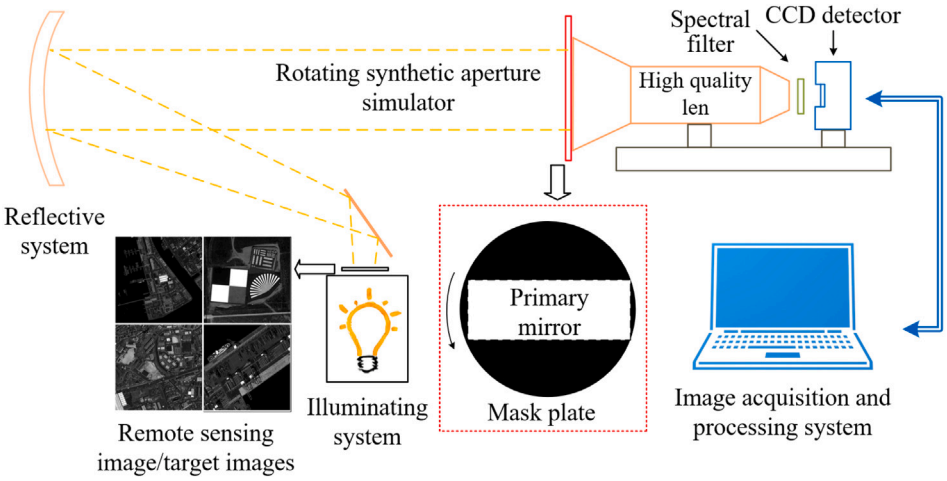

Fig. 12. Design scheme of the imaging experiment platform.

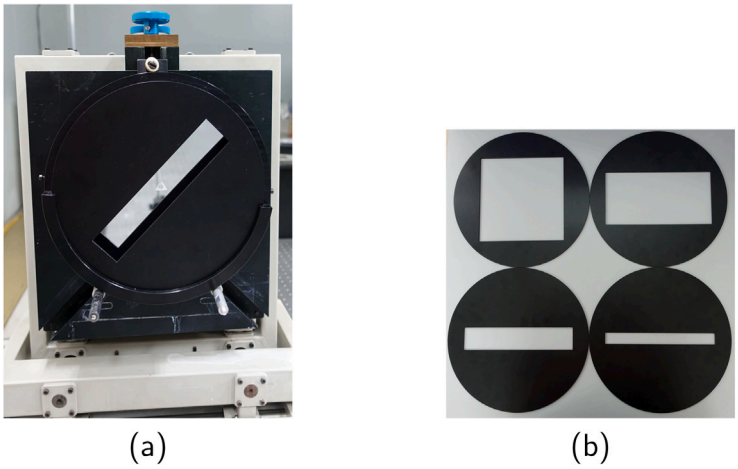

Fig. 13. Experimental components. (a) The primary mirror. (b) The rectangular pupil optical elements.

According to the configuration in SwinIR [52], the RSTB number, STL number, window size, channel number, and attention head number are set to 6, 6, 8, 180, and 6, respectively.

4.2. Experimental results

The proposed SR method is compared with representative and state-of-the-art techniques, including SRGAN, ELAN, Omni-SR, HAT, and RealESRGAN, which are explicit methods utilizing external datasets. Additionally, DualSR, a representative explicit method relying on internal image statistics, and FSSR, a representative implicit method, are included in the comparison. For the HRRSD dataset, the quantitative assessment of these methods using two quality metrics, image structural similarity (SSIM) [58] as well as peak signal-to-noise ratio (PSNR), is presented in Table 2 and Fig. 15. Bicubic interpolation results are also included in the table for comparison. For the test images obtained from the WV3 satellite, Table 3 showcases SR results for six different scenes,

Fig. 14. The dataset used in the experiments, comprised the following scenes: airport, harbor, residential, yard, farmland and forest.

Table 2
SR results for images from HRRSD. The unit of PSNR is decibel (dB). The best and second-best results are indicated in red and blue, respectively.

Dataset	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
HRRSD	Bicubic	28.20	0.8149	27.22	0.7779	26.41	0.7751	25.61	0.7517	25.21	0.7635	24.63	0.7469
	SRGAN	32.14	0.8993	31.12	0.8872	30.23	0.8693	29.19	0.8437	28.76	0.8360	26.23	0.8271
	ELAN	35.19	0.9553	33.20	0.9212	31.71	0.9048	30.08	0.8843	29.04	0.8635	27.92	0.8481
	Omni-SR	34.50	0.9442	32.75	0.9158	31.35	0.8979	29.89	0.8771	28.97	0.8584	27.94	0.8474
	HAT	35.26	0.9561	33.34	0.9239	31.79	0.9066	30.11	0.8862	28.98	0.8601	27.89	0.8456
	Real-ESRGAN	31.27	0.8830	30.32	0.8595	29.55	0.8410	28.50	0.8213	27.78	0.8063	27.21	0.7805
	DualSR	34.06	0.9482	32.01	0.9169	30.85	0.8968	29.65	0.8864	28.89	0.8632	28.01	0.8418
	FSSR	32.16	0.9134	30.64	0.8934	29.96	0.8651	29.08	0.8630	28.47	0.8564	27.97	0.8400
	Proposed	36.60	0.9693	34.66	0.9334	32.59	0.9160	30.62	0.8896	29.21	0.8665	28.01	0.8528

each with six various aspect ratios of the rectangular primary mirror. Furthermore, for a comprehensive overview, Table 4 and Fig. 16 present average results for all test images from WV3.

The self-attention mechanism in Transformer enables the model to more effectively leverage the internal information of images themselves. In the case of the RSA system, this refers to high-resolution information in different directions. Consequently, as can be seen from the above quantitative evaluation results, for the HRRSD dataset, demonstrated superior performance across all six primary mirror aspect ratios, as evidenced by the highest scores achieved in both SSIM and PSNR metrics. Specifically, when the aspect ratio is 3, the SSIM and PSNR of our SR outputs reach 0.9693 and 36.60 dB, respectively, resulting in a 1.38% improvement in SSIM and a 3.80% improvement in PSNR over the second-best method, HAT. For the WV3 dataset, encompassing a total of 36 sets of digital simulation test images with six scenes and six aspect ratios, our proposed method surpasses other approaches in terms of the SSIM metric in 35 sets, as detailed in Tables 3. In terms of the PSNR metric, our method exhibits superior performance in 33 sets of test images. Especially in scenes with rich texture information and high repetition rates, such as ports, residential areas, yards, and forests, our method achieves the best performance under aspect ratios of 3 to 8. Although DualSR outperforms our method in certain scenes based on PSNR metrics, the proposed method achieves significantly better average results for both PSNR and SSIM metrics. This is due to its consideration of the characteristics of the RSA system, unlike DualSR, which solely relies on the similarity of internal image patches. Therefore, our method exhibits robustness to varying aspect ratios of the primary mirror. For different aspect ratios of the primary mirror, when the aspect ratio is 3, the SSIM and PSNR of our SR results reach 0.9714 and 37.49 dB, respectively, which is an 0.81% improvement

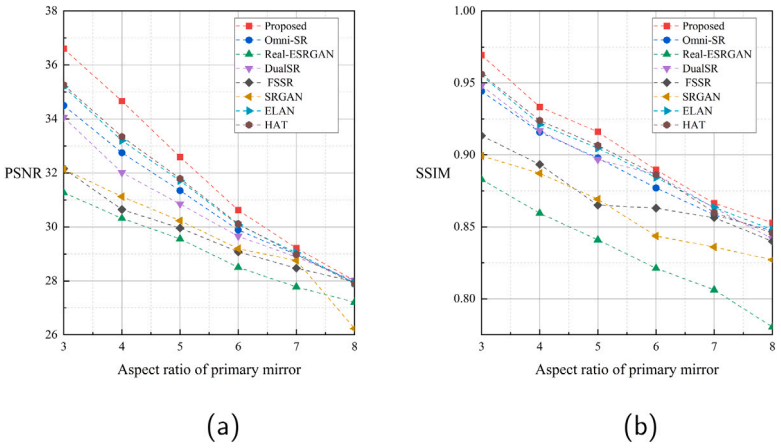

Fig. 15. SR results for images from HRRSD. (a) PSNR. (b) SSIM.

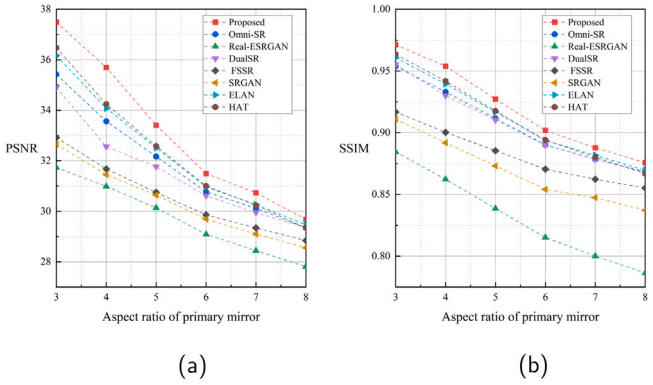

Fig. 16. Average results for images from WorldView-3. (a) PSNR. (b) SSIM.

in SSIM and a 2.77% improvement in PSNR over HAT. When the aspect ratio of the primary mirror is 4 or 5, the proposed method obtains the best results and it is significantly higher than the other methods. Even when the aspect ratio is 6, the SSIM and PSNR of our SR results can still be greater than 0.9 and 31.5 dB, respectively, due to the targeted design of the proposed method based on the image degradation mechanism. When the aspect ratio is large (7 or greater), the performance of our method decreases. This decline in performance is attributed to a considerable reduction in image quality along the shorter side of the mirror, leading to insufficient information for all SR methods to yield satisfactory outcomes. Nevertheless, it is noteworthy that our method still achieves a performance not inferior to the second-best value in terms of the SSIM metric. Additionally, based on Table 4, our method's performance still exceeds that of the second-best method, ELAN, by approximately 0.7%.

Visual presentations are included alongside quantitative assessments to provide qualitative evaluations. Specifically, examples are taken from both the HRRSD dataset and WV3 scenes (as shown in Fig. 6) under the condition of the mirror with an aspect ratio of 4. For the primary mirror rotation angle of 0 degrees, the local enlargement images are presented in Figs. 17(a), 18(a), and 19(a). The SR results for SRGAN, ELAN, Omni-SR, HAT, real-ESRGAN, DualSR, FSSR, and our method are shown in Figs. 17(b)–(i), 18(b)–(i), and 19(b)–(i), respectively. Additionally, when the primary mirror rotation angle is 90 degrees, the local enlargement images are displayed in Figs. 20(a) and 21(a). The SR results for the same methods are presented in Figs. 20(b)–(i) and 21(b)–(i), respectively. The semi-physical imaging experimental images with the primary mirror aspect ratio of 3 and rotation angle of 90 are displayed in Fig. 22, where the processed results are depicted. Specifically, the original local enlargement image is shown in Fig. 22(a), while the SR results using SRGAN, ELAN, Omni-SR, HAT, real-ESRGAN, DualSR, FSSR, and our method are shown in Fig. 22(b)–(i), respectively.

As seen from the visual results above, each method tends to emphasize specific visual characteristics in the SR results, 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 underperform on objective evaluation metrics. On the other hand, the remaining methods tend to produce sharper edges. Nevertheless, the image quality along the shorter side is notably reduced. These generic SR methods are predominantly based on CNNs, making it challenging to leverage long-distance dependencies and self-similarity in remote sensing images. Additionally, the design of these methods often remains independent of imaging system characteristics. Consequently, while some details can be restored, their SR outcomes may still fall short of meeting the resolution requirements of interpretation applications, especially for resolution targets beyond the vertical direction illustrated in Fig. 22. While the SR output obtained by FSSR (Fig. 22(h)) appears to achieve slightly better contrast in the target of the vertical direction, the proposed method can introduce more high-frequency information in other directions to recover the target, which would have been almost completely blurred in the SR results of other methods.

In scenes characterized by a high frequency of visual repetition, such as fields and parking lots, our method demonstrates a more robust capability to recover high-frequency information. As depicted in Fig. 23, in the parking lot scene with a primary mirror aspect ratio of 6 and a rotation angle of 135, the proposed method outperforms ELAN. ELAN is quantitatively evaluated as the second-best in this scene by objective metrics. The proposed method achieves this by reconstructing sharper edges on the lines of parking spots and recovering high-frequency details, which are nearly absent in low-resolution images. Furthermore, in the field scene displayed in Fig. 24, with a primary mirror aspect ratio of 5 and a rotation angle of 45, real-ESRGAN achieves a notable improvement in clarity. However, it erroneously amplifies the spacing between the grass in the field and distorts various shapes in the original image, resulting in undesirable visual artifacts. In contrast, the proposed method utilizes a self-supervised learning approach based on wavelet fusion. This approach yields more natural and reliable SR results while effectively avoiding the generation of "hallucinations" or "artifacts", as illustrated in Figs. 23(d) and 24(d).

4.3. Ablation study

The results of ablating the shifted window mechanism are presented in Table 5. In comparison to self-attention modules employing regular

Table 3
SR results for images from WorldView-3. The unit of PSNR is decibel (dB). The best and second-best results are indicated in red and blue, respectively.

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
Airport	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
	ELAN	35.36	0.9581	33.96	0.9361	31.96	0.9107	30.47	0.8870	29.52	0.8739	28.35	0.8580
	Omni-SR	34.66	0.9508	33.29	0.9300	31.60	0.9053	30.23	0.8821	29.46	0.8709	28.16	0.8556
	HAT	35.74	0.9613	34.15	0.9384	32.01	0.9112	30.55	0.8877	29.50	0.8716	28.22	0.8559
	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	36.59	0.9706	35.22	0.9518	32.86	0.9221	30.95	0.8940	29.80	0.8809	28.18	0.8631
Harbor	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
	ELAN	35.78	0.9710	33.37	0.9523	31.44	0.9282	30.04	0.9161	29.41	0.9050	28.53	0.8945
	Omni-SR	34.80	0.9647	32.66	0.9462	30.97	0.9258	29.77	0.9133	29.11	0.9034	28.33	0.8938
	HAT	36.16	0.9741	33.56	0.9552	31.52	0.9290	30.09	0.9166	29.39	0.9024	28.41	0.8927
	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	37.09	0.9769	35.09	0.9633	32.44	0.9389	30.90	0.9240	30.00	0.9128	29.04	0.9021
Residential	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
	ELAN	35.32	0.9531	33.99	0.9340	31.66	0.9001	30.19	0.8633	29.37	0.8469	28.55	0.8306
	Omni-SR	34.56	0.9439	33.36	0.9230	31.25	0.8906	29.95	0.8591	29.25	0.8434	28.42	0.8280
	HAT	35.63	0.9557	34.16	0.9361	31.74	0.9010	30.23	0.8639	29.35	0.8457	28.44	0.8288
	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	36.77	0.9676	35.39	0.9485	32.48	0.9105	30.82	0.8757	29.96	0.8571	28.80	0.8393
Yard	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
	ELAN	34.44	0.9544	30.91	0.9276	30.52	0.9073	29.07	0.8848	28.57	0.8790	27.93	0.8672
	Omni-SR	33.56	0.9479	30.63	0.9228	30.16	0.9033	28.79	0.8813	28.37	0.8747	27.76	0.8647
	HAT	34.71	0.9572	31.06	0.9300	30.59	0.9081	29.09	0.8851	28.54	0.8769	27.81	0.8636
	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	36.24	0.9692	34.04	0.9485	31.62	0.9206	29.70	0.8930	29.17	0.8852	28.38	0.8738
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
	ELAN	38.50	0.9723	36.67	0.9657	35.48	0.9565	33.61	0.9449	33.27	0.9409	32.32	0.9322
	Omni-SR	38.14	0.9721	36.40	0.9635	35.36	0.9558	33.62	0.9467	33.23	0.9404	32.34	0.9346
	HAT	38.83	0.9754	36.84	0.9682	35.56	0.9574	33.61	0.9442	33.23	0.9402	32.19	0.9281
	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	39.57	0.9794	37.62	0.9719	36.26	0.9634	33.64	0.9522	33.67	0.9422	32.16	0.9351
Forest	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
	ELAN	37.61	0.9562	35.58	0.9215	34.00	0.8985	32.40	0.8661	31.46	0.8445	31.11	0.8354
	Omni-SR	36.82	0.9437	35.04	0.9130	33.67	0.8889	32.23	0.8589	31.31	0.8383	31.04	0.8317
	HAT	37.80	0.9581	35.71	0.9233	34.08	0.8993	32.43	0.8667	31.45	0.8442	30.99	0.8317
	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	38.67	0.9647	36.82	0.9393	34.81	0.9077	32.96	0.8723	31.81	0.8484	31.50	0.8409

Table 4
Average results for images from WorldView-3. The unit of PSNR is decibel (dB). The best and second-best results are indicated in red and blue, respectively.

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
	ELAN	36.17	0.9608	34.08	0.9395	32.51	0.9169	30.96	0.8937	30.27	0.8817	29.46	0.8696
	Omni-SR	35.42	0.9538	33.56	0.9331	32.17	0.9116	30.77	0.8902	30.12	0.8785	29.35	0.8681
	HAT	36.48	0.9636	34.25	0.9419	32.58	0.9177	31.00	0.8940	30.24	0.8802	29.34	0.8668
	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	37.49	0.9714	35.70	0.9539	33.41	0.9272	31.50	0.9019	30.73	0.8878	29.68	0.8757

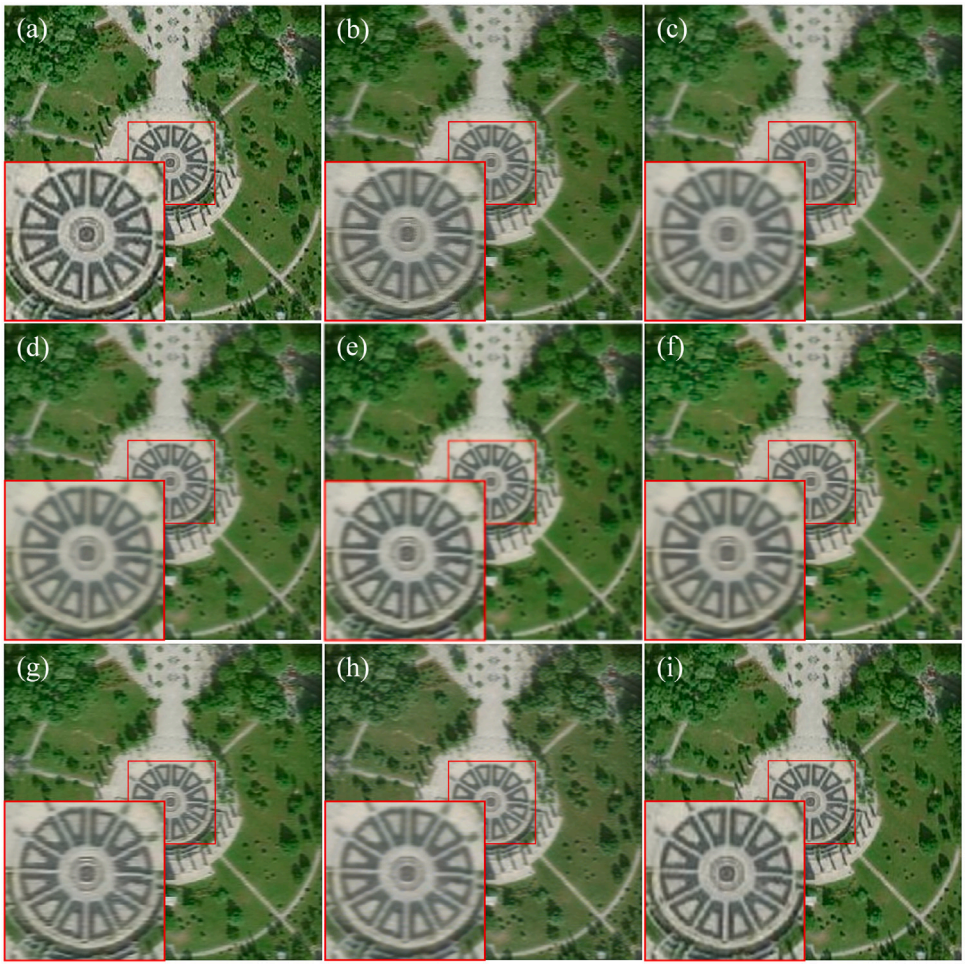

Fig. 17. HR and SR results of the test scene from HRRSD with the rotation angle 0° and the aspect ratio 4. (a) HR. (b) SRGAN. (c) ELAN. (d) Omni-SR. (e) HAT. (f) Real-ESRGAN. (g) DualSR. (h) FSSR. (i) Proposed method.

Table 5
Ablation study on the shifted window mechanism.

	HRRSD		WorldView-3	
	PSNR	SSIM	PSNR	SSIM
w/o shifting	31.52	0.8921	32.57	0.9063
shifted windows	31.95	0.9046	33.09	0.9197

Table 6
Ablation study on the attention mask module.

	HRRSD		WorldView-3	
	PSNR	SSIM	PSNR	SSIM
masked	31.81	0.8999	32.91	0.9144
w/o masking	31.95	0.9046	33.09	0.9197

window partitioning without shifting, the SR network with shifted window partitioning demonstrates superior performance. Specifically, for all test images with aspect ratios ranging from 3 to 8, the average results on the HRRSD dataset show improvements of 0.0125 and 0.43 dB in terms of SSIM and PSNR indices, respectively. Similarly, on the WV3 dataset, it demonstrates enhancements of 0.0134 and 0.52 dB in SSIM and PSNR indices, respectively. These experimental findings highlight the efficacy of utilizing shifted windows to establish connections among windows in preceding layers, thereby contributing to the enhanced performance of remote sensing image SR.

Ablations of the attention mask in the shifted window-based self-attention module are presented in Table 6. The outcomes indicate that eliminating the mask module enhances the utilization of patch recurrence in remote sensing images, facilitating the capture of stronger long-range dependencies inherent in such images. On the HRRSD dataset, the removal of the attention mask led to improvements of 0.0047 and 0.14 dB in SSIM and PSNR, respectively. Similarly, on the WV3 dataset, SSIM and PSNR improved by 0.0053 and 0.18 dB, respectively.

5. Conclusion

In this paper, we propose a self-supervised remote sensing image SR method based on Swin Transformer for the RSA system. By utilizing self-supervision, we leverage the spatial correlations between degraded images at various rotation directions of the rectangular pupil to achieve improved image recovery results while minimizing the risk of "hallucinations". Swin Transformer's content-based interactions between attention weights and image content, along with its shifted window mechanism, can capture stronger long-range dependencies in remote sensing images. Extensive digital simulation and semi-physical imaging experiments compare the proposed SR method with several representative and state-of-the-art techniques, such as SRGAN, RealESRGAN, ELAN, Omni-SR, HAT, DualSR, and FSSR. These experiments use the public dataset HRRSD and images from the WorldView3 satellite, involving six aspect ratios of the primary mirror. For the HRRSD dataset, the proposed method achieved the best performance for all aspect ratios because it fully considers the image degradation mechanism. For the

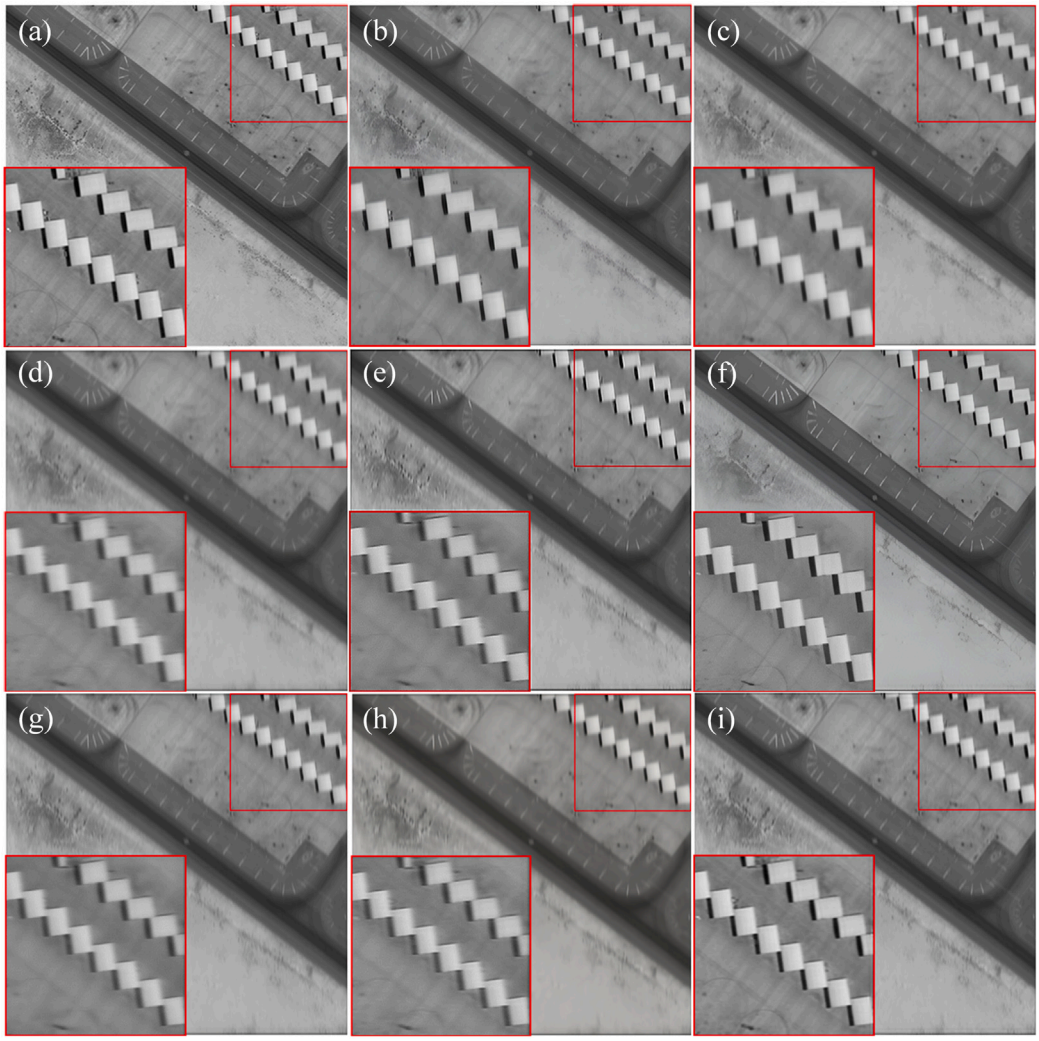

Fig. 18. HR and SR results of the test scene yard with the rotation angle 0° and the aspect ratio 4. (a) HR. (b) SRGAN. (c) ELAN. (d) Omni-SR. (e) HAT. (f) Real-ESRGAN. (g) DualSR. (h) FSSR. (i) Proposed method.

WV3 dataset, encompassing a total of 36 sets of digital simulation test images, our method surpasses other approaches in terms of the SSIM metric in 35 sets and in terms of the PSNR metric in 33 sets. The experimental results also demonstrate the robustness of the proposed method to varying aspect ratios of the primary mirror. It demonstrates significant improvement over other methods for primary mirrors with aspect ratios of 3 to 6. For primary mirrors with larger aspect ratios (7 or greater), the proposed method's performance is slightly degraded but still ranks no lower than the second-best result. Moreover, for scenes with rich texture and high repetition rates, such as residential areas, yards, and forests, the proposed method demonstrates a stronger capability to recover high-frequency information, which can not only obtain the best performance in all aspect ratios but also suppress the generation of artifacts. In future research, we will explore the integration of more advanced Transformer-based models into the SR module. This endeavor aims to enhance the scientific foundation and provide a valuable reference for implementing the RSA imaging technology.

CRediT authorship contribution statement

Yu Sun: Writing – review & editing, Visualization, Validation, Methodology, Conceptualization. **Xiyang Zhi:** Project administration, Investigation, Funding acquisition. **Shikai Jiang:** Software, Formal analysis, Data curation. **Guanghua Fan:** Writing – review & editing, Visualization, Data curation. **Tianjun Shi:** Investigation. **Xu Yan:** Writing – original draft, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data underlying the results presented in this paper are available in the HRRSD dataset.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC) [grant numbers 62305086, 62101160, and 61975043].

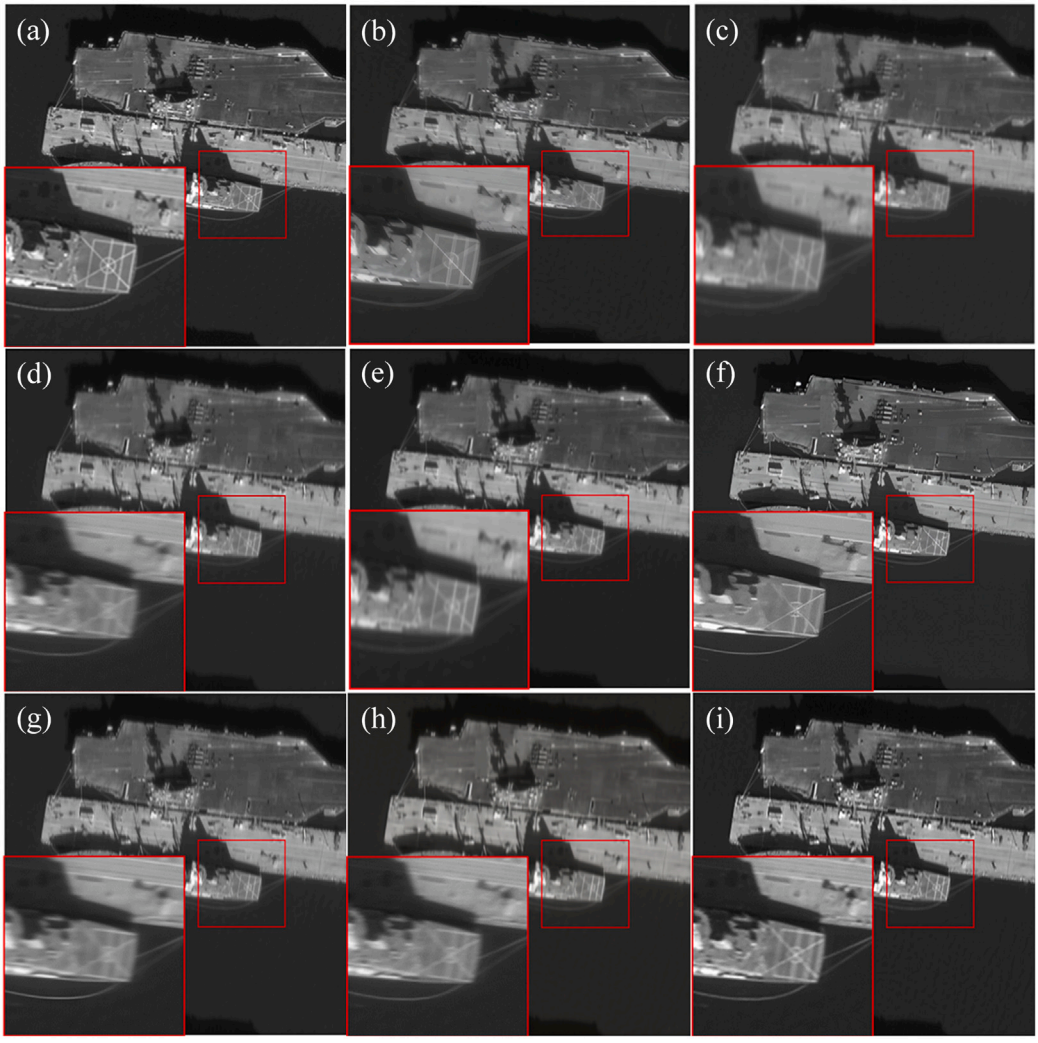

Fig. 19. HR and SR results of the test scene harbor with the rotation angle 0° and the aspect ratio 4. (a) HR. (b) SRGAN. (c) ELAN. (d) Omni-SR. (e) HAT. (f) Real-ESRGAN. (g) DualSR. (h) FSSR. (i) Proposed method.

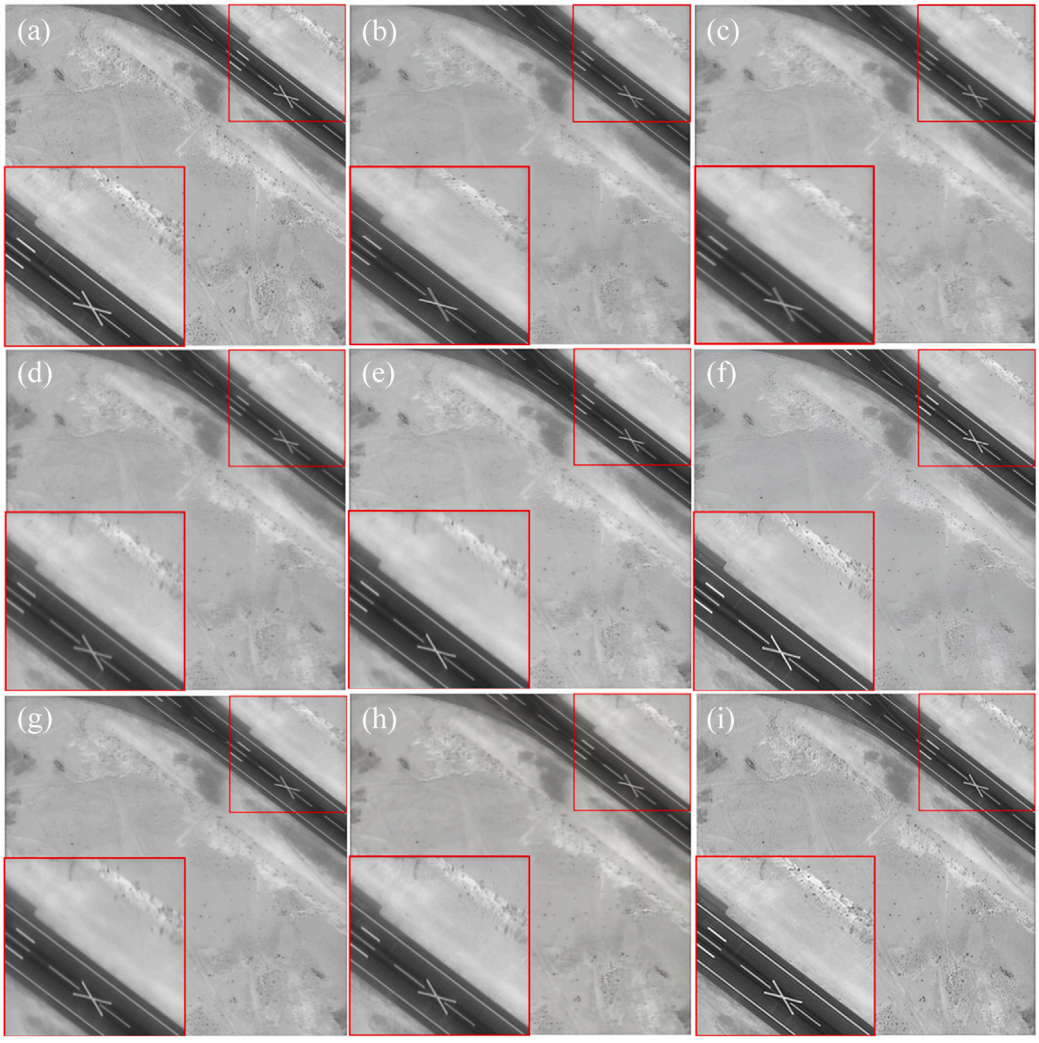

Fig. 20. HR and SR results of the test scene road with the rotation angle 90° and the aspect ratio 4. (a) HR. (b) SRGAN. (c) ELAN. (d) Omni-SR. (e) HAT. (f) Real-ESRGAN. (g) DualSR. (h) FSSR. (i) Proposed method.

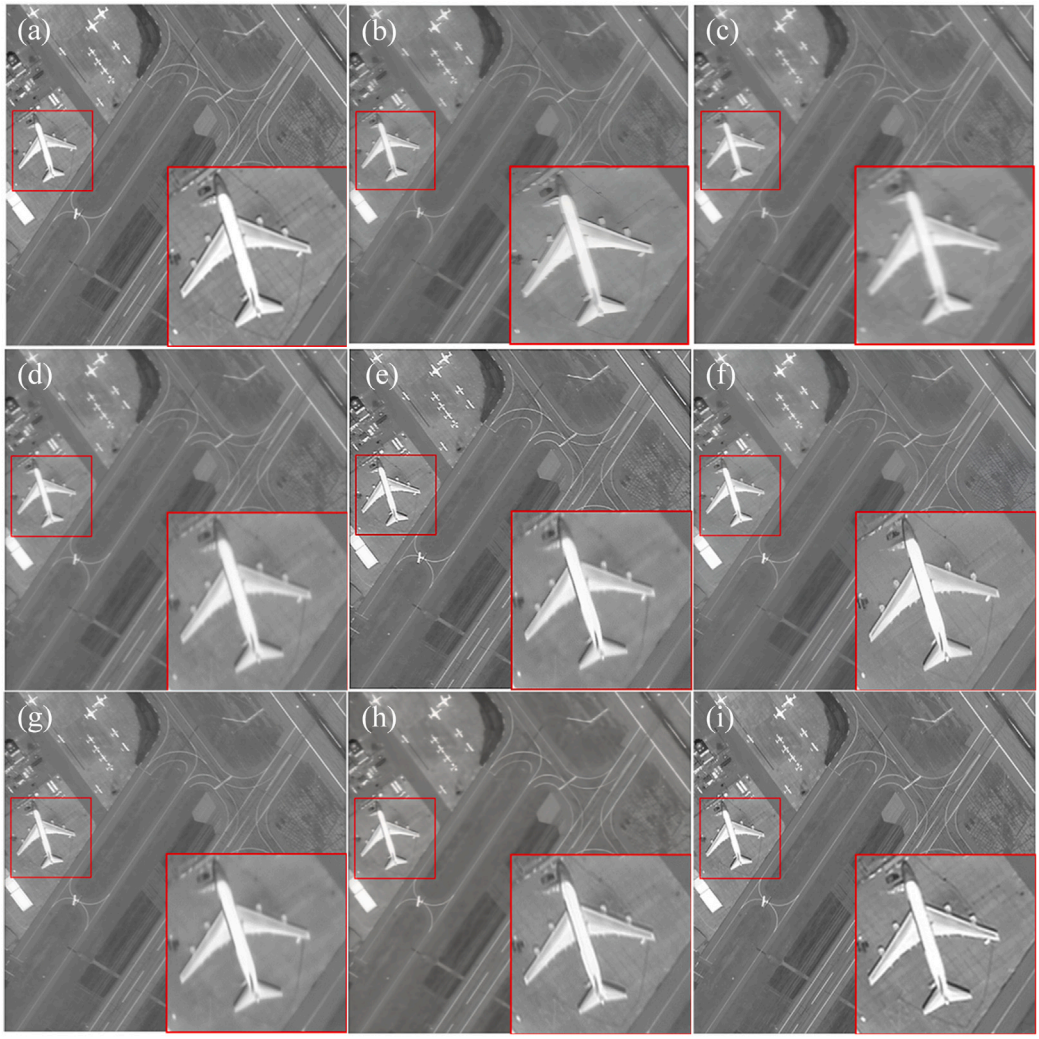

Fig. 21. HR and SR results of the test scene airport with the rotation angle 90° and the aspect ratio 4. (a) HR. (b) SRGAN. (c) ELAN. (d) Omni-SR. (e) HAT. (f) Real-ESRGAN. (g) DualSR. (h) FSSR. (i) Proposed method.

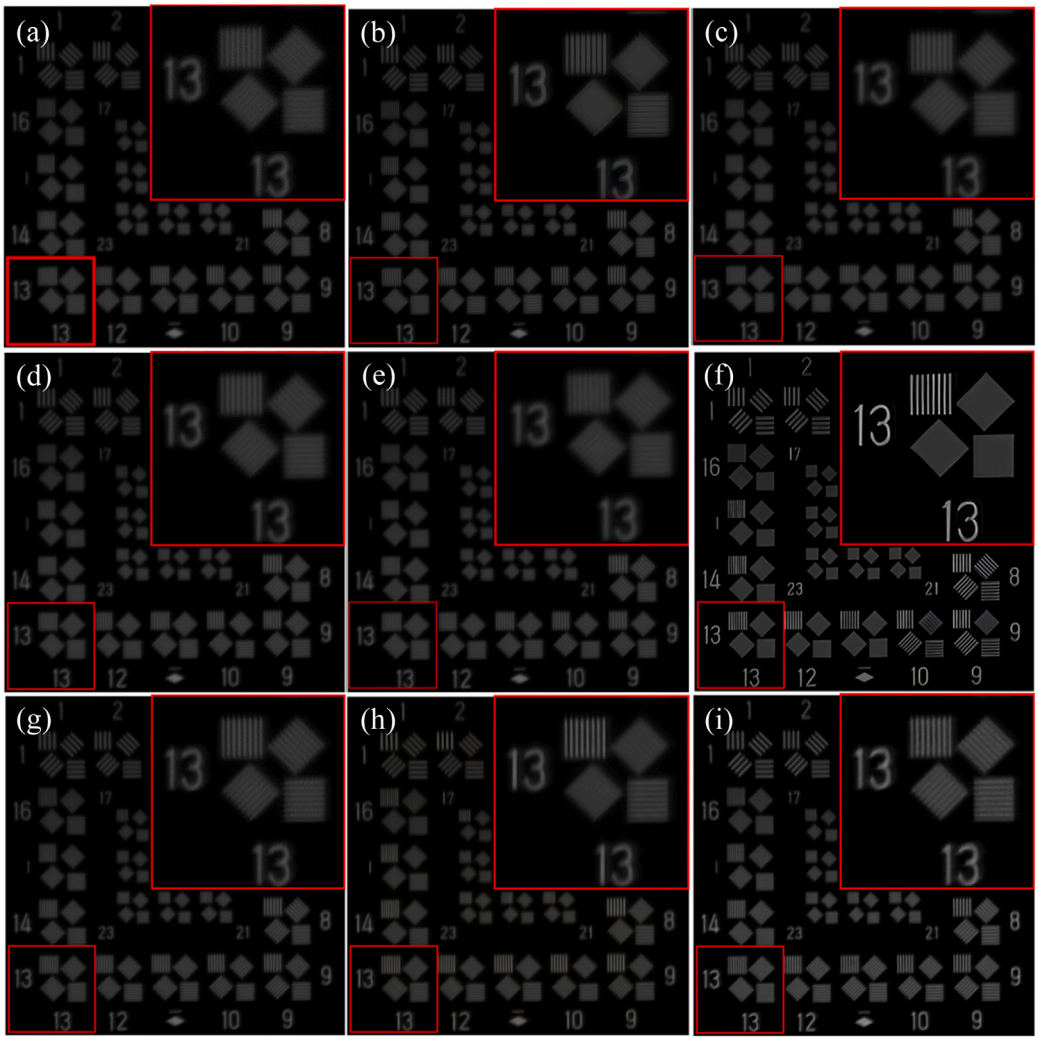

Fig. 22. SR results of the semi-physical experimental image with the rotation angle 90° and the aspect ratio 3. (a) LR. (b) SRGAN. (c) ELAN. (d) Omni-SR. (e) HAT. (f) Real-ESRGAN. (g) DualSR. (h) FSSR. (i) Proposed method.

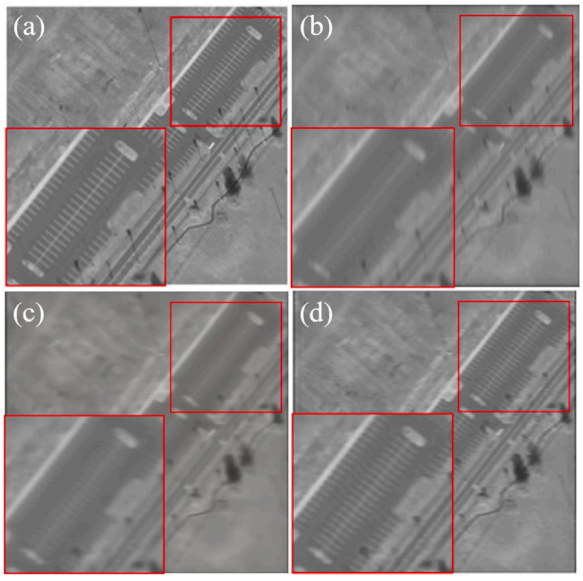

Fig. 23. SR results of the test image parking lot with the rotation angle 135° and the aspect ratio 6. (a) HR. (b) LR. (c) ELAN. (d) Proposed method.

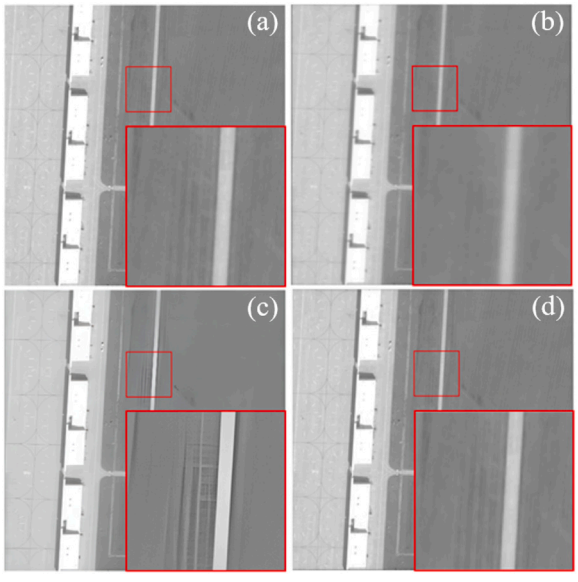

Fig. 24. SR results of the test image field with the rotation angle 45° and the aspect ratio 5. (a) HR. (b) LR. (c) Real-ESRGAN. (d) Proposed method.

References

[1] J. Guo, J. Zhao, L. Zhu, D. Gong, Status and trends of the large aperture space optical remote sensor, in: 2018 IEEE International Conference on Mechatronics and Automation, ICMA, IEEE, 2018, pp. 1861–1866.

[2] F. Bao, Y. Li, J. Gao, Carbonaceous aerosols remote sensing from geostationary satellite observation, Part I: Algorithm development using critical reflectance, Remote Sens. Environ. 287 (2023) 113459.

[3] H. Zhang, H. Liu, W. Xu, Z. Lu, Large aperture diffractive optical telescope: A review, Opt. Laser Technol. 130 (2020) 106356.

[4] M.R. Rai, J. Rosen, Optical incoherent synthetic aperture imaging by superposition of phase-shifted optical transfer functions, Opt. Lett. 46 (7) (2021) 1712–1715.

[5] J. Wu, F. Yang, L. Cao, Resolution enhancement of long-range imaging with sparse apertures, Opt. Lasers Eng. 155 (2022) 107068.

[6] J. Tang, K. Wang, Z. Ren, W. Zhang, X. Wu, J. Di, G. Liu, J. Zhao, RestoreNet: a deep learning framework for image restoration in optical synthetic aperture imaging system, Opt. Lasers Eng. 139 (2021) 106463.

[7] W. Zhao, X. Zhang, J. Wang, Y. Gu, An end-to-end deep convolutional neural network for image restoration of sparse aperture imaging system in geostationary orbit, in: Optoelectronic Imaging and Multimedia Technology IX, Vol. 12317, SPIE, 2023, pp. 193–201.

[8] S. Jiang, X. Zhi, Y. Dong, W. Zhang, D. Wang, Inversion restoration for space diffractive membrane imaging system, Opt. Lasers Eng. 125 (2020) 105863.

[9] S. Jiang, X. Zhi, W. Zhang, D. Wang, J. Hu, C. Tian, Global information transmission model-based multiobjective image inversion restoration method for space diffractive membrane imaging systems, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–12.

[10] R. Niu, S. Jiang, X. Zhi, J. Hu, W. Zhang, J. Gong, Development and analysis of space-based diffractive optical imaging techniques, in: 2023 International Conference for Advancement in Technology, ICONAT, IEEE, 2023, pp. 1–6.

[11] X. Zhi, S. Jiang, L. Zhang, D. Wang, J. Hu, J. Gong, Imaging mechanism and degradation characteristic analysis of novel rotating synthetic aperture system, Opt. Lasers Eng. 139 (2021) 106500.

[12] X. Zhi, S. Jiang, L. Zhang, J. Hu, L. Yu, X. Song, J. Gong, Multi-frame image restoration method for novel rotating synthetic aperture imaging system, Results Phys. 23 (2021) 103991.

[13] T. Geng, X.-Y. Liu, X. Wang, G. Sun, Deep shearlet residual learning network for single image super-resolution, IEEE Trans. Image Process. 30 (2021) 4129–4142.

[14] B. Rasti, Y. Chang, E. Dalsasso, L. Denis, P. Ghamisi, Image restoration for remote sensing: Overview and toolbox, IEEE Geosci. Remote Sens. Mag. (2021).

[15] R. Ryan, B. Baldridge, R.A. Schowengerdt, T. Choi, D.L. Helder, S. Blonski, IKONOS spatial resolution and image interpretability characterization, Remote Sens. Environ. 88 (1–2) (2003) 37–52.

[16] F. De Lussy, P. Kubik, D. Greslou, V. Pascal, P. Gigord, J.P. Cantou, PLEIADES-HR image system products and quality-PLEIADES-HR image system products and geometric accuracy, in: Proceedings International Society for Photogrammetry and Remote Sensing International Conference, 2005, pp. 17–20.

[17] A. Liu, Y. Liu, J. Gu, Y. Qiao, C. Dong, Blind image super-resolution: A survey and beyond, IEEE Trans. Pattern Anal. Mach. Intell. (2022).

[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, 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, 2017, pp. 4681–4690.

[19] X. Zhang, H. Zeng, S. Guo, L. Zhang, Efficient long-range attention network for image super-resolution, in: European Conference on Computer Vision, Springer, 2022, pp. 649–667.

[20] H. Wang, X. Chen, B. Ni, Y. Liu, J. Liu, Omni aggregation networks for lightweight image super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22378–22387.

[21] X. Chen, X. Wang, J. Zhou, Y. Qiao, C. Dong, Activating more pixels in image super-resolution transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22367–22377.

[22] X. Wang, L. Xie, C. Dong, Y. Shan, Real-esrgan: Training real-world blind super-resolution with pure synthetic data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1905–1914.

[23] S. Bell-Kligler, A. Shocher, M. Irani, Blind super-resolution kernel estimation using an internal-gan, Adv. Neural Inf. Process. Syst. 32 (2019).

[24] M. Emad, M. Peemen, H. Corporaal, Dualsr: Zero-shot dual learning for real-world super-resolution, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1630–1639.

[25] J. Kim, C. Jung, C. Kim, Dual back-projection-based internal learning for blind super-resolution, IEEE Signal Process. Lett. 27 (2020) 1190–1194.

[26] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, L. Lin, Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 701–710.

[27] M. Fritsche, S. Gu, R. Timofte, Frequency separation for real-world super-resolution, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop, ICCVW, IEEE, 2019, pp. 3599–3608.

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[29] D.C. Lepcha, B. Goyal, A. Dogra, V. Goyal, Image super-resolution: A comprehensive review, recent trends, challenges and applications, Inf. Fusion 91 (2023) 230–260.

[30] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R.E. Sheriff, C. Zhu, Real-world single image super-resolution: A brief review, Inf. Fusion 79 (2022) 124–145.

[31] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12009–12019.

[32] Q. Zhang, Y. Xu, J. Zhang, D. Tao, Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond, Int. J. Comput. Vis. (2023) 1–22.

[33] F.A. Dharejo, I.I. Ganapathi, M. Zawish, B. Alawode, M. Alathbah, N. Werghi, S. Javed, SwinWave-SR: Multi-scale lightweight underwater image super-resolution, Inf. Fusion 103 (2024) 102127.

[34] W. Zhang, W. Zhao, J. Li, P. Zhuang, H. Sun, Y. Xu, C. Li, CVANet: Cascaded visual attention network for single image super-resolution, Neural Netw. 170 (2024) 622–634.

[35] Y. Sun, X. Zhi, S. Jiang, G. Fan, X. Yan, W. Zhang, Image fusion for the novelty rotating synthetic aperture system based on vision transformer, Inf. Fusion 104 (2024) 102163.

[36] X. Ning, Z. Yu, L. Li, W. Li, P. Tiwari, DILF: Differentiable rendering-based multi-view image–language fusion for zero-shot 3D shape understanding, Inf. Fusion 102 (2024) 102033.

[37] S. Wei, H. Cheng, B. Xue, X. Shao, T. Xi, Low-cost and simple optical system based on wavefront coding and deep learning, Appl. Opt. 62 (23) (2023) 6171–6179.

[38] L. Tang, J. Yuan, J. Ma, Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network, Inf. Fusion 82 (2022) 28–42.

[39] M.V. Conde, F. Vasluianu, R. Timofte, BSRAW: Improving blind RAW image super-resolution, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 8500–8510.

[40] C. Wang, X. Ning, L. Sun, L. Zhang, W. Li, X. Bai, Learning discriminative features by covering local geometric space for point cloud analysis, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.

[41] C. Wang, X. Ning, W. Li, X. Bai, X. Gao, 3D person re-identification based on global semantic guidance and local feature aggregation, IEEE Trans. Circuits Syst. Video Technol. (2023).

[42] C. Wang, C. Wang, W. Li, H. Wang, A brief survey on RGB-d semantic segmentation using deep learning, Displays 70 (2021) 102080.

[43] C. Wang, H. Wang, X. Ning, S. Tian, W. Li, 3D point cloud classification method based on dynamic coverage of local area, J. Softw. 34 (4) (2022) 1962–1976.

[44] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, J. Yan, FFCA-yolo for small object detection in remote sensing images, IEEE Trans. Geosci. Remote Sens. (2024).

[45] D. Wang, J. Zhang, B. Du, G.-S. Xia, D. Tao, An empirical study of remote sensing pretraining, IEEE Trans. Geosci. Remote Sens. (2022).

[46] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, L. Zhang, Advancing plain vision transformer toward remote sensing foundation model, IEEE Trans. Geosci. Remote Sens. 61 (2022) 1–15.

[47] S. Lei, Z. Shi, Hybrid-scale self-similarity exploitation for remote sensing image super-resolution, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–10.

[48] S. Lei, Z. Shi, W. Mo, Transformer-based multistage enhancement for remote sensing image super-resolution, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–11.

[49] S. Maeda, Unpaired image super-resolution using pseudo-supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 291–300.

[50] M. Zontak, M. Irani, Internal statistics of a single natural image, in: CVPR 2011, IEEE, 2011, pp. 977–984.

[51] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693.

[52] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1833–1844.

[53] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[54] A. Berk, P. Conforti, R. Kennett, T. Perkins, F. Hawes, J. Van Den Bosch, MODTRAN® 6: A major upgrade of the MODTRAN® radiative transfer code, in: 2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, WHISPERS, IEEE, 2014, pp. 1–4.

[55] A. Berk, P. Conforti, F. Hawes, An accelerated line-by-line option for MODTRAN combining on-the-fly generation of line center absorption within 0.1 cm-1 bins and pre-computed line tails, in: Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXI, Vol. 9472, SPIE, 2015, pp. 405–415.

[56] Y. Sun, X. Zhi, L. Zhang, S. Jiang, T. Shi, N. Wang, J. Gong, Characterization and experimental verification of the rotating synthetic aperture optical imaging system, Sci. Rep. 13 (1) (2023) 17015.

[57] Y. Zhang, Y. Yuan, Y. Feng, X. Lu, Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection, IEEE Trans. Geosci. Remote Sens. 57 (8) (2019) 5535–5548.

[58] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.