Information Fusion 104 (2024) 102163

Contents lists available at ScienceDirect

Information Fusion

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

Image fusion for the novelty rotating synthetic aperture system based on vision transformer

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

[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, United Kingdom

ARTICLE INFO

Keywords:
Image fusion
Vision transformer
Optical remote sensing
Rotating synthetic aperture

ABSTRACT

Rotating synthetic aperture (RSA) technology offers a promising solution for achieving large-aperture and lightweight designs in optical remote-sensing systems. It employs a rectangular primary mirror, resulting in noncircular spatial symmetry in the point-spread function, which changes over time as the mirror rotates. Consequently, it is crucial to employ an appropriate image-fusion method to merge high-resolution information intermittently captured from different directions in the image sequence owing to the rotation of the mirror. However, existing image-fusion methods have struggled to address the unique imaging mechanism of this system and the characteristics of the geostationary orbit in which the system operates. To address this challenge, we model the imaging process of a noncircular rotating pupil and analyse its on-orbit imaging characteristics. Based on this analysis, we propose an image-fusion network based on a vision transformer. This network incorporates inter-frame mutual attention and intra-frame self-attention mechanisms, facilitating more effective extraction of temporal and spatial information from the image sequence. Specifically, mutual attention was used to model the correlation between pixels that were close to each other in the spatial and temporal dimensions, whereas long-range spatial dependencies were captured using intra-frame self-attention in the rotated variable-size attention block. We subsequently enhanced the fusion of spatiotemporal information using video swin transformer blocks. Extensive digital simulations and semi-physical imaging experiments on remote-sensing images obtained from the WorldView-3 satellite demonstrated that our method outperformed both image-fusion methods designed for the RSA system and state-of-the-art general deep learning-based methods.[1]

1. Introduction

Geostationary remote-sensing satellites offer optical remote sensing with both high spatial and temporal resolutions, making them essential components of space-based observation technology [1–4]. Owing to their high orbital altitude, geostationary satellites require larger apertures to ensure imaging quality. Currently, breakthroughs in aperture limitations are primarily achieved through technologies such as segmented mirror [5–7], membrane diffraction imaging [8–11], optical synthetic aperture [12–14], and rotating synthetic aperture (RSA) technology. Among these, the RSA system, which originated from the rotating slit-aperture telescope concept, stands out as a superior alternative [15]. It uses a rotatable primary mirror with a large aspect ratio, as shown in Fig. 1. During the imaging process, rotation of the primary mirror generates a sequence of images containing high-resolution information in different directions [16]. Using image-fusion methods, RSA systems can achieve an imaging quality nearly identical to or even higher than that of equivalent circular-aperture systems [17].

Previous studies proposed several image-fusion techniques specifically designed for RSA systems. Zackay et al. [18–20] were the first to introduce a system-specific nonblind restoration method using matched filters. However, fusion results obtained using these approaches remain unclear. Zhou et al. [21] treated the image fusion of an RSA system as an image deblurring problem and proposed a method using multiframe deconvolution. Similarly, Lv et al. [22] introduced an image-fusion method that restores the Fourier spectrum. Although these methods demonstrate good fusion effects for sequence images with significant noise, they do not fully consider the effects of satellite-platform

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

[1] Dataset is available at https://github.com/sherlockyusun/RSA-Fuse.

https://doi.org/10.1016/j.inffus.2023.102163
Received 9 October 2023; Received in revised form 15 November 2023; Accepted 28 November 2023
Available online 29 November 2023

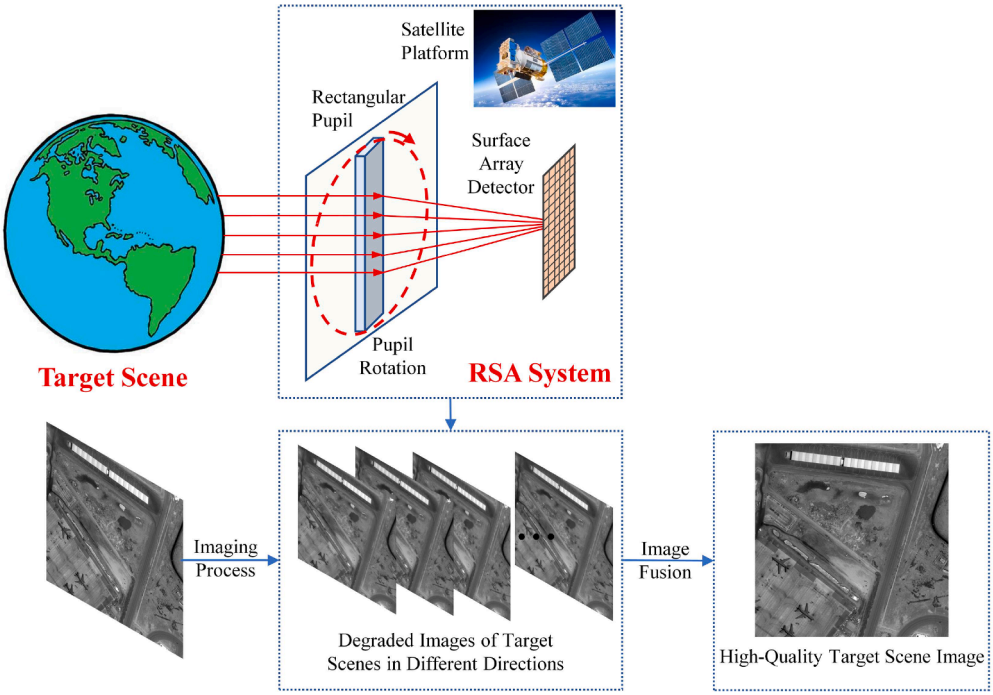

Fig. 1. Imaging process and image fusion of the rotating synthetic aperture (RSA) system.

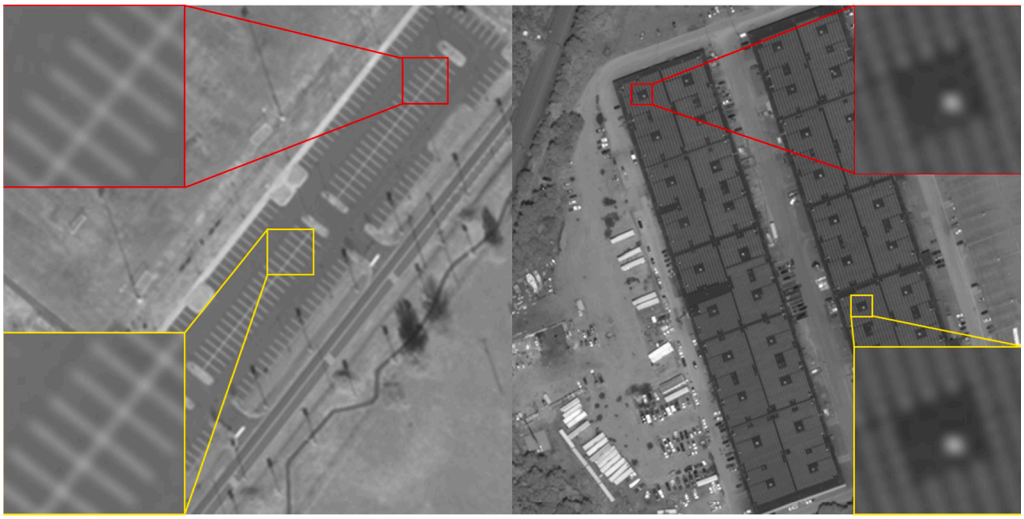

Fig. 2. Long-range dependency in remote-sensing images.

vibration during on-orbit applications of the RSA system. Registration is a typical approach for addressing inter-frame displacement. Zhi et al. [23] proposed a registration method tailored to the imaging mechanism of an RSA system. The registration accuracy of this method depends on the directional gradient prior of the image. Inaccurate motion estimation can have a significant impact on the quality of fusion results [24]. Another limitation of the aforementioned methods is that they do not belong to the category of deep learning-based approaches. Using trainable feature extractors, deep learning methods can mine semantic features and perform better than traditional methods [25–30]. Consequently, deep neural networks have been successfully applied to computer vision, particularly in the field of image fusion [31–37].

More precisely, image fusion in the RSA system aims to merge temporal information from multiple adjacent frames. These sequential frames include inter-frame displacement caused by vibrations in the satellite platform, which often results in misalignment. Therefore, the main challenge lies in correctly handling the inter-frame displacement and fully utilising the additional temporal information. The dynamic imaging method of rotating the pupil causes frames within an image sequence to possess varying resolutions in the same direction, creating difficulties in accurately estimating the motion. Inaccurate motion estimation can lead to unreasonable motion compensation, resulting in the loss of critical prior image information and the introduction of errors and artefacts. However, the continuous rotation of the primary mirror during the imaging process forms a three-dimensional dataset in which each frame in the temporal dimension exhibits a high correlation with its adjacent frames. Hence, inter-frame relations in the temporal dimension are as important as intra-frame relations in the spatial dimension. However, the effective utilisation of information in the temporal domain based on the unique imaging mechanism and on-orbit characteristics of the RSA system remains an unexplored issue that requires further investigation.

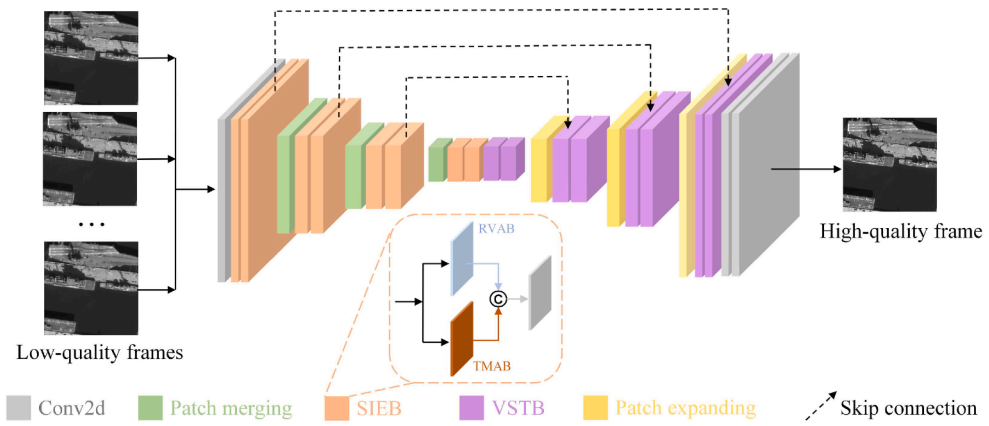

Fig. 3. Overall framework of the proposed network.

Based on the preceding analysis, it is evident that directly applying existing deep learning-based image-fusion methods to the RSA system is inappropriate. This is because these methods fail to consider the imaging mechanism and orbital imaging characteristics of the system. Furthermore, most current deep learning approaches are based on convolutional neural networks (CNNs). Convolutional operations in CNNs primarily capture localised features owing to their inherent local inductive biases. However, remote-sensing images, as shown in Fig. 2, reveal a high correlation between pixels and patches, even at larger spatial distances. This long-range dependency on remote-sensing images can result in highly correlated content lying beyond the receptive field of the convolutional kernel and going unnoticed by the kernel. This renders CNNs incapable of establishing long-range spatial dependencies.

To address this concern, we consider the fact that the self-attention mechanism of vision transformer (ViT) is designed specifically to calculate correlations among pixels within an image. This resonates with the goal of fusing sequential images to preserve high-resolution information in different directions. Consequently, it enables a more efficient utilisation of both intra-frame spatial relations and inter-frame temporal relations in a remote-sensing image sequence. Therefore, we treat the image quality improvement of the RSA system as a multisource visual fusion problem and present an end-to-end network that leverages ViT as its foundation. This is the first deep learning-based image-fusion method to enhance the image quality of an RSA system. As shown in Fig. 3, the network leverages intra-frame self-attention and inter-frame mutual attention to extract and fuse spatiotemporal information. In the spatiotemporal information extraction block (SIEB), mutual attention in the temporal mutual attention block (TMAB) is employed to model the correlation between pixels that are close to each other in the spatial and temporal dimensions, whereas long-range dependencies are captured using self-attention in the rotated varied-size attention block (RVAB) [38]. Specifically, in the RVAB, we utilise rotated varied-size self-attention to extract spatial information within frames. This approach introduced shift, scale, and rotation factors based on window-based attention to capture diverse local windows. By allowing windows of various locations, sizes, shapes, and angles, the model can better address objects with different orientations and scales that are prevalent in remote-sensing images. This capability is advantageous for extracting additional contextual information. Explicit motion compensation operations typically rely on bilinear or bicubic resampling operations. However, the weights of such operations are inaccessible, rendering interpolation irreversible. This can lead to a loss of information [39]. Therefore, we avoided explicit alignment operations and used TMAB directly to extract temporal information. After extracting information from the spatial and temporal domains, we use video swin transformer blocks (VSTBs) [40] to merge the extracted features. The transformer can implicitly establish connections for unaligned pixels by calculating the mutual attention between adjacent frames, thereby adaptively preserving information from different frames. This is equivalent to implicitly performing motion estimation and image warping at the feature level to avoid information loss and artefacts. Finally, we conducted a comparative analysis between our proposed method and image-fusion techniques designed for the RSA system, as well as state-of-the-art general deep learning-based image-fusion methods. The results of both digital simulations and semi-physical imaging experiments validate the effectiveness of our method.

The main contributions of this work can be summarised as follows:

(1) Analysis of temporal periodicity and spatial asymmetry characteristics of rotating rectangular pupils in RSA system.
(2) Establishment of on-orbit dynamic imaging characteristic model for the RSA system.
(3) An image-fusion network based on ViT aligned with the system's imaging mechanism and on-orbit imaging characteristics was proposed. By integrating attention into the time dimension, the network leveraged intra-frame self-attention and inter-frame mutual attention to extract and fuse spatiotemporal information.

The remainder of this paper is structured as follows: Section 2 discusses the imaging mechanism of the RSA system and analyses its imaging characteristics during on-orbit operations. In Section 3, we introduce the details of the proposed image-fusion network. Section 4 validates the effectiveness and accuracy of the proposed method using digital and semi-physical simulation experiments. Finally, Section 5 concludes the study.

2. Imaging characteristics of the rotating synthetic aperture system

Traditional optical remote-sensing systems typically use circular primary mirrors, resulting in circularly symmetric point-spread functions (PSFs) that degrade identically in all directions. In contrast, the RSA system uses a rectangular primary mirror with a high aspect ratio. Specifically, the pupil function of the mirror at time t is

$$P_{rect}(\xi,\eta,t) = \mathrm{rect}\left(\frac{\xi\cos(wt+\varphi_0)-\eta\sin(wt+\varphi_0)}{a}\right) \mathrm{rect}\left(\frac{\xi\sin(wt+\varphi_0)+\eta\cos(wt+\varphi_0)}{b}\right) \quad (1)$$

where a and b are the length and width of the rectangle, respectively; w is the angular velocity of the mirror rotation; and φ_0 is the initial phase.

Applying the Fourier transform to Eq. (1) and taking the modulo square, we obtain the PSF at time t [41]

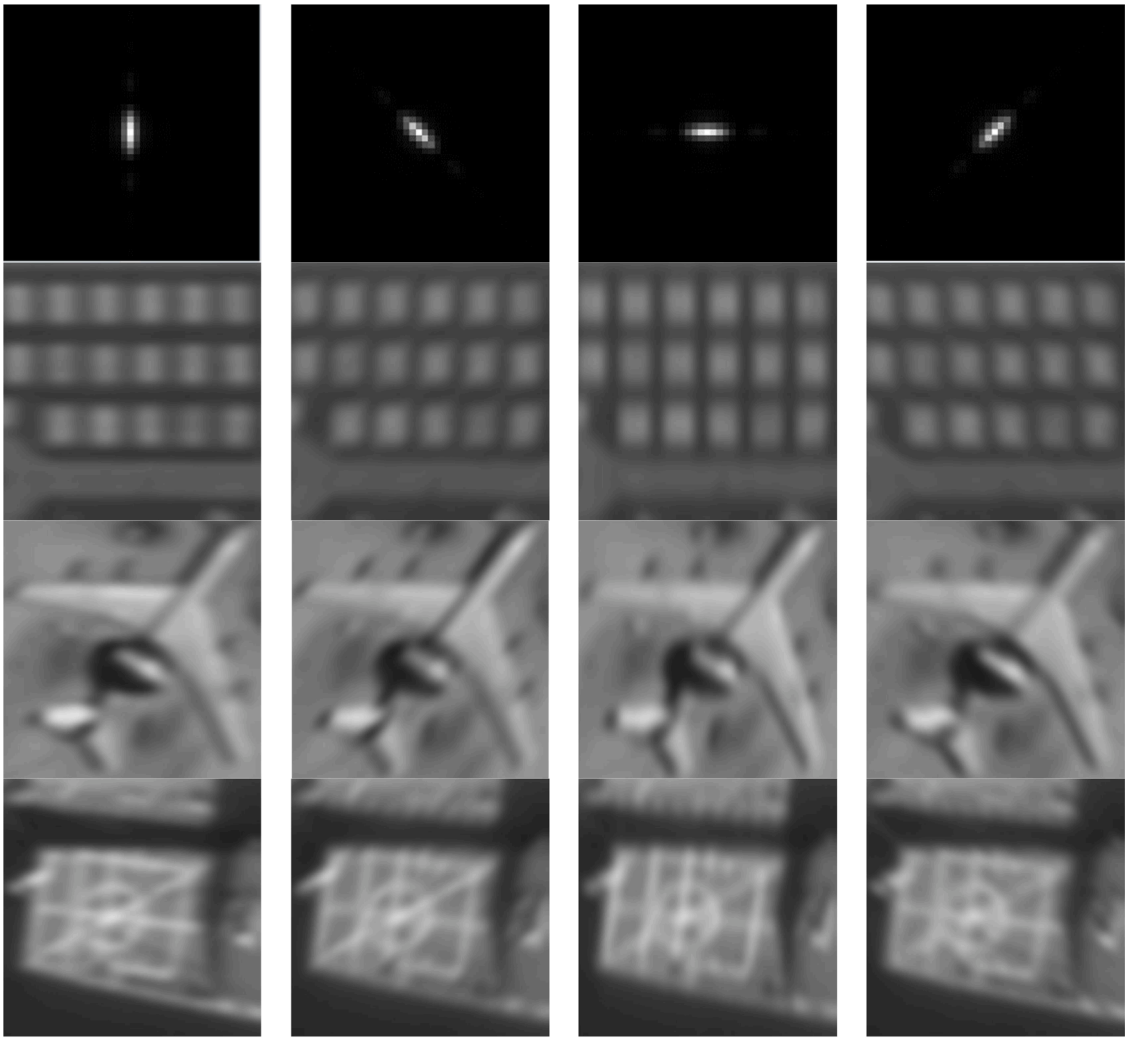

Fig. 4. PSFs and degraded images with various rotation angles. (a) 0 °, (b) 45 °, (c) 90 °, and (d) 135 °

$$PSF(x,y,t) = ab\mathrm{sinc}(a(x\cos(wt+\varphi_0) - y\sin(wt+\varphi_0))) \times \mathrm{sinc}(b(x\sin(wt+\varphi_0) + y\cos(wt+\varphi_0))) \quad (2)$$

Fig. 4 shows the PSF of the primary mirror at various rotation angles, along with the resulting images. These images reveal that the resolution of the acquired images differs significantly in different directions, as evidenced by the edge and texture details of the square buildings, aircraft, and ships.

In addition to the aforementioned unique pupil shape, the complex coupling of links in the imaging process also affects the image quality of the RSA system during in-orbit applications. More specifically, the system's imaging-link commences with the target scene and culminates with the digital image. It encompasses numerous links, including the atmosphere, optical system, on-orbit vibration of the satellite platform, detector photoelectric conversion and sampling, and imaging circuits. The primary factor affecting image fusion is satellite-platform vibration, which causes relative displacement (i.e., image shift) between the object and detector. In particular, the angular rotation of the satellite platform under a single frequency sinusoidal vibration mode has a substantial effect on image quality. The platform's rotation angles along its three spatial axes, namely, pitch, roll, and yaw, are respectively denoted by θ_p, θ_r, and θ_y.

In the satellite coordinate system *xyz* (image plane coordinate system *xy*), the amplitudes of the image shift A_{y1} and A_{x2} resulting from the vibrations of the satellite in the roll and pitch axes, respectively, are

$$\begin{aligned} A_{y1} &= f\tan\theta_r \\ A_{x2} &= f\tan\theta_p \end{aligned} \quad (3)$$

where f is the focal length.

The image shift amplitudes generated on the x and y axes owing to the vibration in the yaw axis direction can be represented as

$$\begin{aligned} A_{x3} &= d\sin\theta_y \\ A_{y3} &= d\left(1-\cos\theta_y\right) \end{aligned} \quad (4)$$

where d represents the pixel size of the detector.

In the case of harmonic oscillation in the cosine form, the image shift at time t can be represented by the image shift functions $A_x(t)$ and $A_y(t)$ in the satellite coordinate system as follows:

$$\begin{aligned} A_x(t) &= A_{x2}\cos(w_{vib}t+\psi_2) + A_{x3}\cos(w_{vib}t+\psi_3) \\ A_y(t) &= A_{y1}\cos(w_{vib}t+\psi_1) + A_{y3}\cos(w_{vib}t+\psi_3) \end{aligned} \quad (5)$$

where ψ_1, ψ_2, and ψ_3 denote the initial phase angles of the axis vibrations during system exposure and w_{vib} represents the frequency of the satellite-platform vibration.

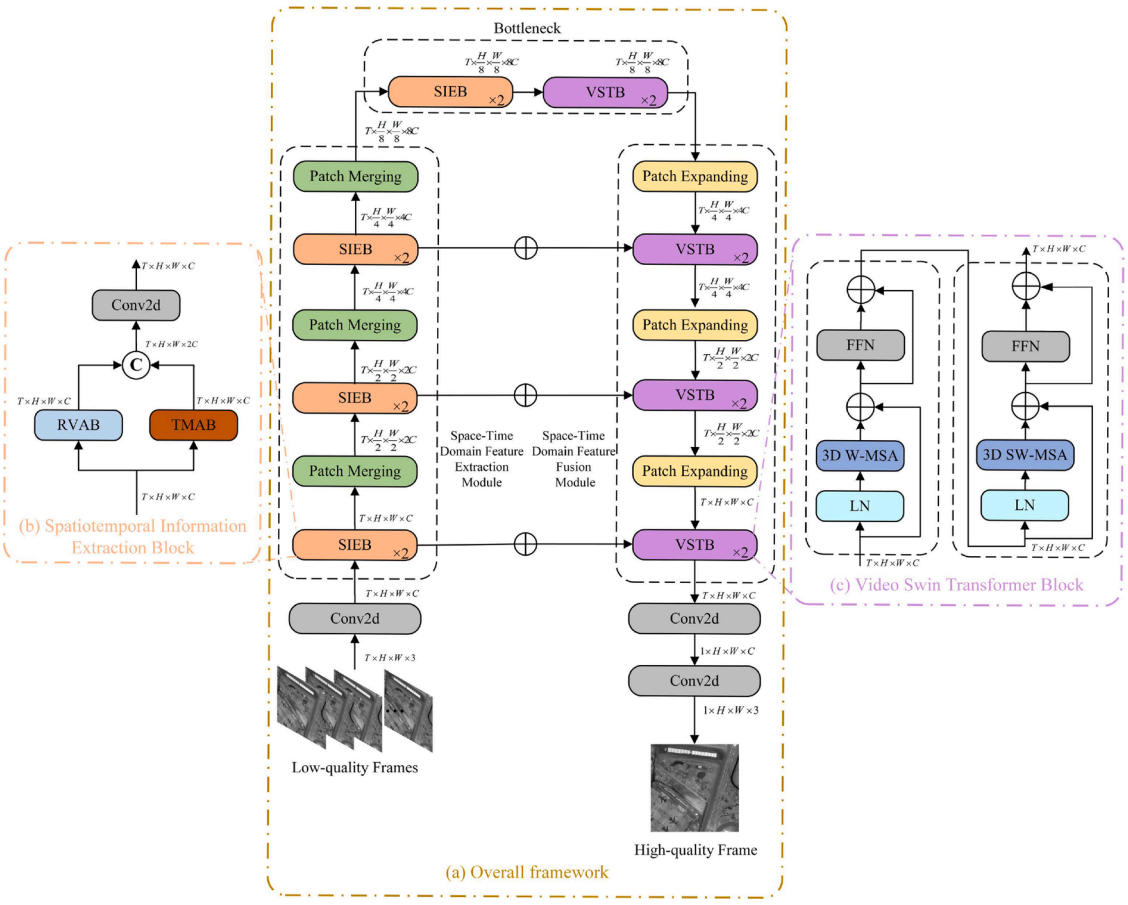

Fig. 5. Architecture of the proposed image-fusion network.

3. Methods

3.1. Overall architecture of the proposed network

The aim of the proposed image-fusion network is to create one high-quality frame $H \in \mathbb{R}^{1\times H\times W\times 3}$ from T low-quality frames $I \in \mathbb{R}^{T\times H\times W\times 3}$. Fig. 5 illustrates the network architecture, which essentially adopts a U-shaped structure. The input is a sequence of low-quality frames, where T, H, and W represent the sequence length, height, and width of the low-quality frames, respectively. First, the network utilised a two dimensional (2D) convolutional layer to map I into tokens $\mathbf{X}_0 \in \mathbb{R}^{T\times H\times W\times C}$, where C denotes the channel number. Next, $\mathbf{X}_0$is processed through the space–time domain feature-extraction module consisting of SIEBs and patch merging layers to generate hierarchical features. The patch merging layer doubles the channels and downsamples the feature maps, whereas $\mathrm{X}_i \in \mathbb{R}^{T\times\frac{H}{2^i}\times\frac{W}{2^i}\times 2^i C}$ denotes the tokens passing through the i_{th} patch merging layer. Then, X_i passes through one SIEB and one VSTB. These two blocks act as the bottleneck in the U-shaped network. Analogously, we employ some VSTBs and patch-expanding layers as the space–time domain feature fusion module. During the fusion of spatio-temporal information, the features are gradually upsampled, and finally, the features are restored to the original size $T \times H \times W \times C$. We use skip connections to mitigate information loss caused by patch merging and the burden of feature learning. At the end of the network, there are two 2D convolutional layers. Specifically, after passing through the last VSTB in the space–time domain feature fusion module, we merge the information of each channel of the features $\mathbf{X}_0^{'} \in \mathbb{R}^{T\times H\times W\times C}$ in the time dimension. For C channels, $\mathbf{X}_0^{'}$ can be regarded as C feature maps of size $H \times W \times T$. To merge these maps, we employ C 1×1 2D spatial convolution kernels, resulting in C feature maps of size $H \times W \times 1$. After merging, the size of the features becomes $1 \times H \times W \times C$. Finally, a 2D convolutional layer is utilised to adjust the number of output channels, generating a high-quality frame H.

3.2. Space–time domain feature-extraction module

To better capture self-similarity within a single frame and across adjacent frames, we introduce content-based interactions between attention weights and image content in the process of space–time domain feature extraction. However, it is worth noting that remote-sensing images often have large sizes, and directly applying global self-attention will result in quadratic computational complexity with respect to the token number. Thus, we employed a window-based attention approach in the space–time domain feature-extraction module to replace the global attention, reducing the computational complexity to a linear correlation with the image size.

The space–time domain feature-extraction module is composed of several SIEBs and patch merging layers. As illustrated in Fig. 5(b), each SIEB consists of an RVAB, TMAB, and a 2D convolutional layer.

3.2.1. Window-based attention

Both RVABs and TMABs within the space–time domain feature-extraction module utilise window-based attention. Window-based attention [42] differs from the standard global self-attention in its local attention and window-transfer mechanisms. Specifically, for an input of size $T \times H \times W \times C$, denoted as $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$, window-based attention initially reshapes the input by partitioning it into non-overlapping $M \times M$ local windows, denoted as $\mathrm{X} \in \mathbb{R}^{T\times\frac{H}{M}\times\frac{W}{M}\times M^2\times C}$, where $\frac{HW}{M^2}$ is the total number of windows. For each window, the input features are represented as $\mathrm{X}_w \in \mathbb{R}^{T\times M^2\times C}$, thus, all input features can be

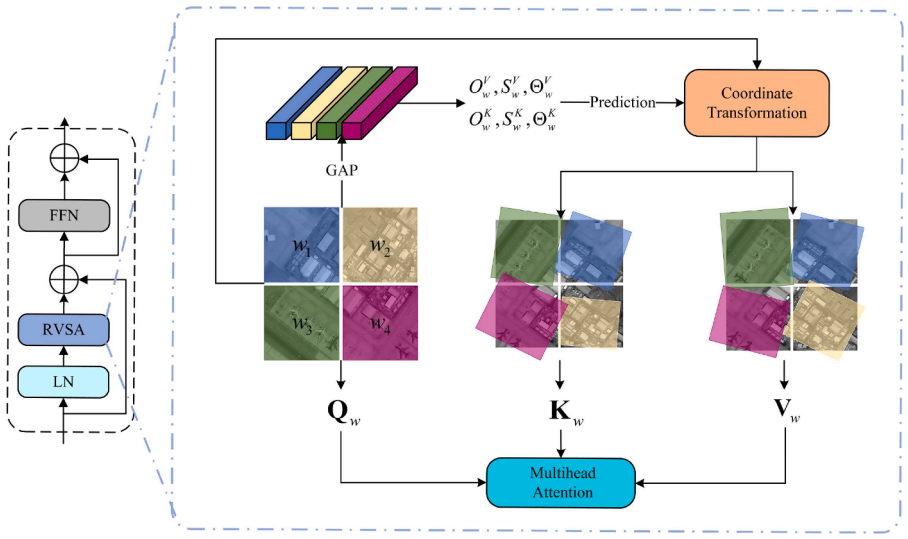

Fig. 6. Illustration of the rotated varied-size attention block.

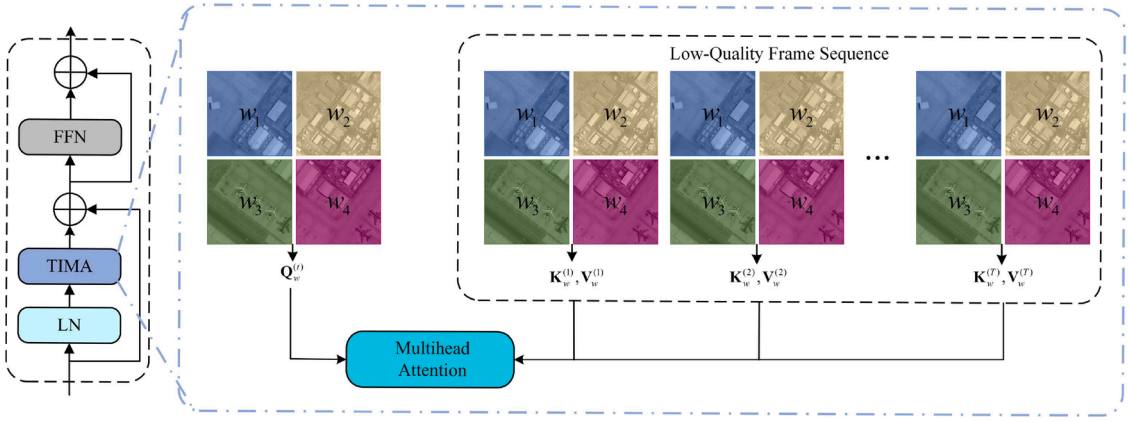

Fig. 7. Illustration of the temporal mutual attention block.

denoted as $\left\{X_{w_i} | i = 1, \cdots, \frac{HW}{M^2}\right\}$. Subsequently, the standard multihead self-attention was calculated for each window. Let h denote the number of heads, and the *query, key* and *value* matrices are represented by $\{\mathbf{Q}_{w_i}^{(j)}\}$, $\{\mathbf{K}_{w_i}^{(j)}\}$, and $\{\mathbf{V}_{w_i}^{(j)}\}$, respectively. Here, i indexes the window ($i = 1, \cdots, \frac{HW}{M^2}$), and j indexes the head ($j = 1, \cdots, h$).

The attention calculations are then conducted in each non-overlapping local window:

$$\mathbf{Z}_{w_i}^{(j)} = \text{Attention}\left(\mathbf{Q}_{w_i}^{(j)}, \mathbf{K}_{w_i}^{(j)}, \mathbf{V}_{w_i}^{(j)}\right) = \text{softmax}\left(\frac{\mathbf{Q}_{w_i}^{(j)}\left(\mathbf{K}_{w_i}^{(j)}\right)^{'}}{\sqrt{C}}\right)\mathbf{V}_{w_i}^{(j)} \tag{6}$$

where $\mathbf{Q}_{w_i}^{(j)}, \mathbf{K}_{w_i}^{(j)}, \mathbf{V}_{w_i}^{(j)}, \mathbf{Z}_{w_i}^{(j)} \in \mathbb{R}^{T \times M^2 \times C^{'}}$ and $C^{'} = \frac{C}{h}$

Finally, the features $\left\{\mathbf{Z}_{w_i}^{(j)} | i = 1, \cdots, \frac{HW}{M^2}, j = 1, \cdots, h\right\}$ are concatenated to restore the original shape of the input. Specifically, features from diverse non-overlapping windows are concatenated along the spatial dimension, whereas features from diverse heads are concatenated along the channel dimension.

3.2.2. Rotated varied-size attention block

As is well known, in remote-sensing images, it is a common challenge to deal with a variety of objects that can be oriented in different ways and come in various sizes. However, the window size in the original window-based attention operation was fixed, and the window was always horizontal or vertical. Let (x_c, y_c), (x_{ul}, y_{ul}), and (x_{lr}, y_{lr}) denote the coordinates of the pixels at the centre, upper-left corner, and lower-right corner of the window, respectively.

$$\begin{bmatrix} x_{ul} \\ y_{ul} \\ x_{lr} \\ y_{lr} \end{bmatrix} = \begin{bmatrix} x_c \\ y_c \\ x_c \\ y_c \end{bmatrix} + \begin{bmatrix} x_l^r \\ y_l^r \\ x_r^r \\ y_r^r \end{bmatrix} \tag{7}$$

where x_l^r, y_l^r, x_r^r and y_r^r denote the distances between the coordinates of the corner and centre points, respectively.

A fixed and unchangeable window is clearly not conducive to extracting the contextual information of targets with different orientations and scales in remote-sensing images. To address this problem, we introduce additional learnable parameters and implement rotated varied-size attention blocks in the space–time domain feature-extraction module. In accordance with Section 3.2.1, we independently calculated the intra-frame self-attention for the input features of each time step (i.e. every low-quality frame) to extract the spatial domain information within the frame more effectively.

As illustrated in Fig. 6, the RVAB comprises layer normalisation, rotated varied-size multihead attention (RVSA), and feed-forward network (FFN) [38]. Unlike the original window-based attention operations, RVAB does not consider fixed-size window partitions in a fixed

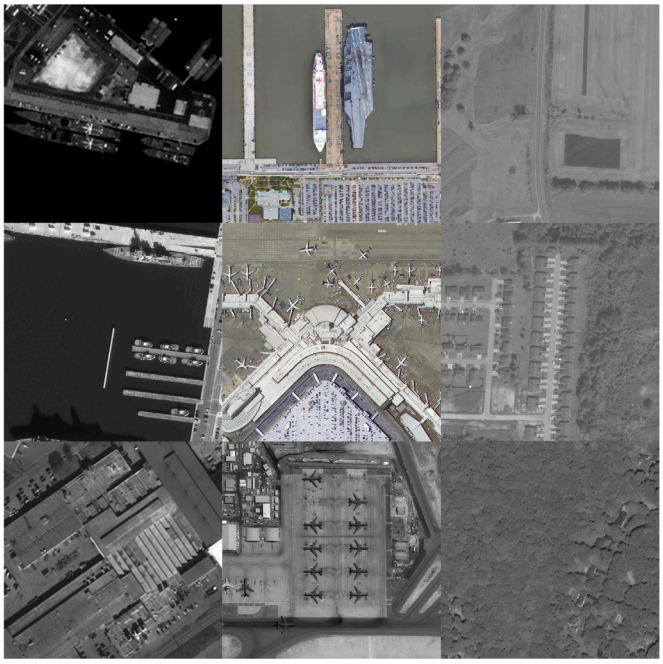

Fig. 8. Target scenes from WorldView3. The ground resolution of these remote-sensing images is approximately 0.4 m.

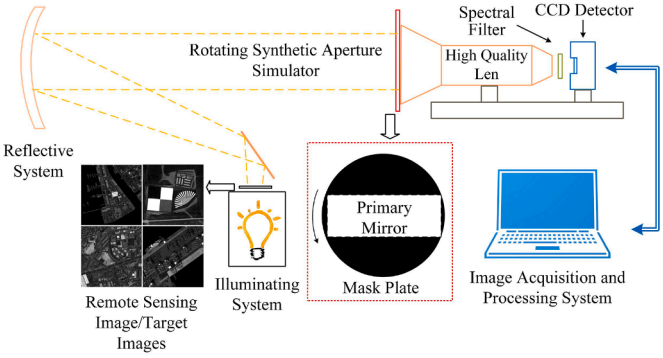

Fig. 9. Design scheme of the imaging experiment platform.

orientation. Instead, it generates windows with varying positions, sizes, shapes, and angles through learnable shift, scale, and rotation parameters, that is O_w^K, S_w^K, and Θ_w, respectively. Specifically, for each window, separate prediction layers can be used to predict the shift, scale, and rotation parameters of *key* and *value* tokens based on the input features.

$$O_w^K, S_w^K, \Theta_w^K = \mathrm{Linear}_K(\mathrm{LeakyReLU}(\mathrm{GAP}(\mathbf{X}_w))) \tag{8}$$

$$O_w^V, S_w^V, \Theta_w^V = \mathrm{Linear}_V(\mathrm{LeakyReLU}(\mathrm{GAP}(\mathbf{X}_w))) \tag{9}$$

where GAP denotes the global average pooling operation.

Thereafter, based on the aforementioned parameters, the initial window is transformed, and the transformed coordinates of the corner points $(x'_{l/r}, y'_{l/r})$ are calculated as follows:

$$\begin{bmatrix} x'_{l/r} \\ y'_{l/r} \end{bmatrix} = \begin{bmatrix} x^c \\ y^c \end{bmatrix} + \begin{bmatrix} o_x \\ o_y \end{bmatrix} + \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x^r_{l/r} \cdot s_x \\ y^r_{l/r} \cdot s_y \end{bmatrix} \tag{10}$$

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

The *key* and *value* features were then sampled from the transformed windows and utilised to calculate multihead self-attention [38]. The remaining steps are similar to those outlined in Section 3.2.1 and need not be reiterated at this point. More importantly, different heads can produce windows with varying positions, sizes, and shapes. This implies that the RVAB is better suited for extracting information from multiple target objects of various scales and orientations.

3.2.3. Temporal mutual attention block

The rotation of the rectangular pupil and coupling of various factors during the imaging process can lead to unequal information content in adjacent frames. Therefore, the inter-frame temporal relation also plays a crucial role in the image fusion of the RSA system. To this end, we propose a TMAB by introducing the calculation of attention into the time dimension and leveraging the transformer's powerful modelling capabilities to capture self-similarity among pixels over time.

Because of the high temporal resolution of geostationary orbit satellites, inter-frame displacement in RSA system images occurs at the subpixel level. Specifically, as per the analysis in [16], for an RSA system with a 90 m effective focal length, the geometric distortion is usually no more than 0.2 pixels. The experiments in [39,43] indicate that alignment operations only positively impact pixels with a large motion (inter-frame displacement greater than five). This means that traditional explicit registration and motion compensation methods are unsuitable

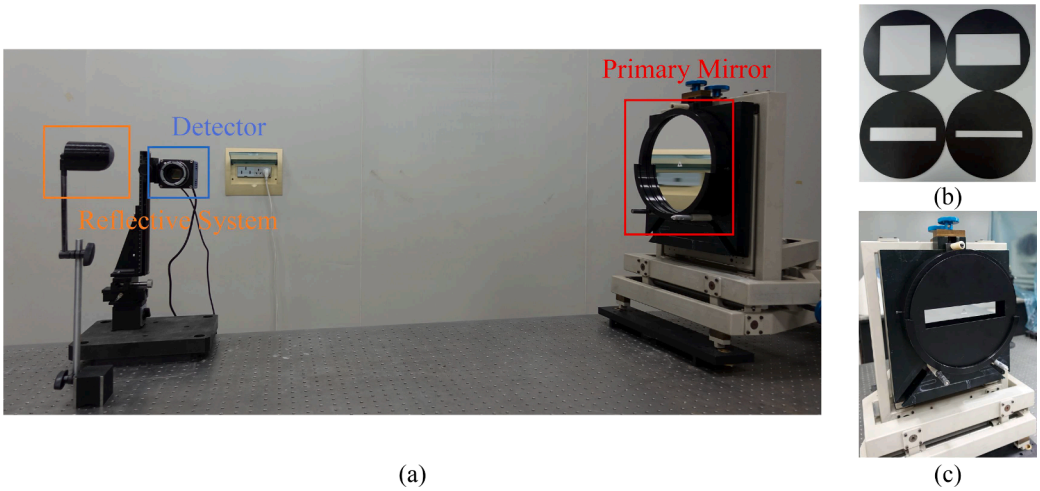

Fig. 10. (a) Semi-physical imaging experiment platform, (b) rectangular pupil optical elements, and (c) the primary mirror with an element is placed in front.

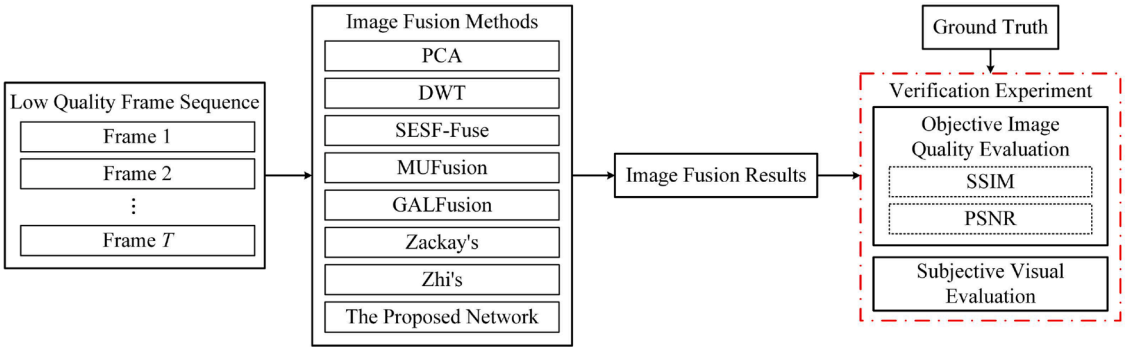

Fig. 11. Experimental flowchart.

for extracting time-domain information because inaccurate motion estimation and image warping can cause error accumulation. Inaccurate motion estimation is a combination of true inter-frame displacement and random errors, and warping operations based on random errors can destroy subpixel-level information [39]. Moreover, misregistration and misalignment in the preorder stage can negatively affect information extraction in the subsequent stage, leading to error propagation that affects the subsequent fusion module and results in artefacts in the fused high-quality frames. To address this issue, we replaced the alignment module with a temporal mutual attention approach based on a window-based transformer. For sub-pixel-level inter-frame displacement, a window-based transformer can implicitly establish connections for unaligned pixels because it lacks a local inductive bias, similar to CNNs. This dynamic aggregation of adjacent frames at the subpixel level is necessary for the efficient fusion of time-domain information. The mutual attention we employed can adaptively detect correlations between pixels in different frames and align them implicitly at the feature level, which is beneficial for extracting additional information and facilitating fusion. Furthermore, it helps to avoid the occurrence of black hole artefacts in explicit motion compensation operations owing to the lack of matching positions.

TMAB comprises layer normalisation, temporal inter-frame mutual attention (TIMA), and FFN, as shown in Fig. 7. The calculation of TIMA is similar to that of window-based intra-frame self-attention in RVAB; however, instead of generating different locations, sizes, shapes, and angles of windows, each head in TMAB focuses on high-resolution information preserved in different directions in the images obtained at different times. Specifically, for each window, $\mathbf{X}_{w_i}^{(t)} \in \mathbb{R}^{M^2 \times C}$ and $\mathbf{Q}_{w_i}^{(t)} \in \mathbb{R}^{M^2 \times \frac{C}{T}}$ represent the input vector and *query* matrices at time t in the sequence, respectively. In the multihead attention calculation of TMAB, the *key* and *value* matrices $\{\mathbf{K}_{w_i}^{(p)}\}$ and $\{\mathbf{V}_{w_i}^{(p)}\}$ are calculated based on the input features $\{\mathbf{X}_{w_i}^{(p)}\}$ of the window at the same position in each frame. Here, $\mathbf{K}_{w_i}^{(p)}, \mathbf{V}_{w_i}^{(p)} \in \mathbb{R}^{M^2 \times \frac{C}{T}}$, i indexes the window ($i = 1, \cdots, \frac{HW}{M^2}$), and p indexes the head, that is, the frame number ($p = 1, \cdots, T$).

Subsequently, mutual attention is calculated as:

$$\mathbf{Z}_{w_i}^{(t,p)} = \text{Attention}\left(\mathbf{Q}_{w_i}^{(t)}, \mathbf{K}_{w_i}^{(p)}, \mathbf{V}_{w_i}^{(p)}\right) = \text{softmax}\left(\frac{\mathbf{Q}_{w_i}^{(t)}\left(\mathbf{K}_{w_i}^{(p)}\right)'}{\sqrt{\frac{C}{T-1}}}\right)\mathbf{V}_{w_i}^{(p)} \in \mathbb{R}^{M^2 \times \frac{C}{T}}, p = 1, \cdots, T \tag{11}$$

Thereafter, the features $\mathbf{Z}_{w_i}^{(t,p)} \in \mathbb{R}^{M^2 \times \frac{C}{T}}$ are concatenated along the channel dimension to generate the features $\mathbf{Z}_{w_i}^{(t)} \in \mathbb{R}^{M^2 \times C}$ at time t.

Table 1
Quantitative results of PSNR (dB). The best result is in red while the second-best result is in blue.

Method	Scene type					
	Airport	Farmland	Forest	Harbour	Residential	Average
DWT	26.14	29.26	27.91	27.29	26.25	27.37
PCA	25.82	28.43	26.93	27.09	25.90	26.83
Zackay's [19]	23.81	27.08	25.65	25.18	24.22	25.19
Zhi's [23]	30.23	32.14	31.68	30.65	29.93	30.93
GALFusion [35]	27.72	31.23	29.27	28.49	27.58	28.86
SESF-Fuse [32]	27.91	31.55	29.68	28.65	27.84	29.12
MUFusion [34]	30.73	33.63	32.18	30.83	30.23	31.52
Proposed	31.83	35.70	33.27	31.88	31.15	32.77

Because $\mathbf{K}_w$and $\mathbf{V}_w$ are generated from adjacent frames, $\mathbf{Z}_{w_i}^{(t,p)}$ reflects the correlation between elements in the adjacent frames. By calculating inter-frame mutual attention, TMAB can explore similar features in adjacent frames during the process of extracting information along the time dimension for the subsequent fusion module. Finally, to restore the shape of input features, $\mathbf{Z}_{w_i}^{(t)}$ are concatenated along the time dimension.

3.3. Space–time domain feature fusion module

In our space–time domain feature-extraction module, we employed a combination of spatial self-attention and temporal mutual attention mechanisms. This approach enables us to approximate global self-attention by utilising local self-attention in the subsequent space–time domain feature fusion module. Following the U-shaped network architecture, this module comprises the same number of VSTBs and patch-expanding layers as the SIEBs and patch merging layers in the space–time domain feature-extraction module. The VSTB (shown in Fig. 5(c)) replaces the multihead self-attention in the standard global transformer with either three dimensional (3D) window-based multihead self-attention or 3D shifted window-based multihead self-attention. It consists of layer normalisation, 3D multihead self-attention, and an FFN. Therefore, we always use two consecutive VSTBs in the space–time domain feature fusion module, with one calculating the 3D window-based multihead self-attention and the other calculating the 3D shifted window-based multihead self-attention. Finally, two 2D convolutional layers follow the space–time domain feature fusion module to adjust the channel dimensions and generate a high-quality output frame.

4. Experiments

4.1. Experimental setup

4.1.1. Datasets

Digital and semi-physical imaging simulation experiments were conducted separately to verify the efficacy of the proposed method. The digital simulation experiment utilised high-resolution remote-sensing images and employed the simulation method in [41] to simulate the imaging quality degradation process of the RSA system, resulting in a low-resolution version of the high-resolution remote-sensing images for the construction of the dataset. The input image comprised various scenes, including airports, farmlands, forests, harbours, and residential areas. These images were derived from WorldView-3 satellite data and were downloaded from the official website of Maxar Technologies Inc. (https://www.maxar.com/). Some of the target scenes are shown in Fig. 8. To further substantiate the superiority of the proposed method in addressing satellite-platform vibrations, we referred to the analysis presented in [16] and deliberately selected parameters that induced a more pronounced degradation in image quality. To be precise, for an optical system with an effective focal length of 90 m, we set the vibration frequency of the satellite platform to 20 Hz, the angular amplitude to 0.05 μrad, and the rotational angular velocity of the primary mirror to 0.1 rad/s.

According to [22], to reconstruct high-quality fusion images, it is

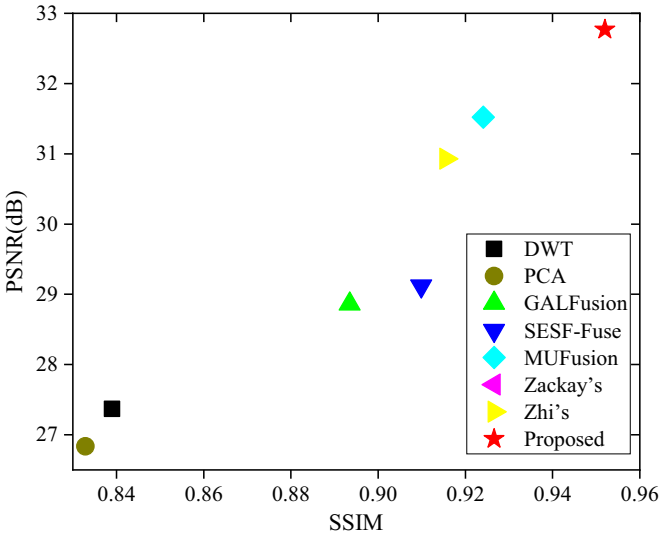

Fig. 12. Quantitative comparison of two evaluation metrics among different methods.

Table 2
Quantitative results of SSIM. The best result is in red while the second-best result is in blue.

Method	Scene type					
	Airport	Farmland	Forest	Harbour	Residential	Average
DWT	0.8239	0.8670	0.8528	0.8301	0.8213	0.8390
PCA	0.8173	0.8571	0.8422	0.8328	0.8150	0.8329
Zackay's [19]	0.8912	0.9201	0.9067	0.8951	0.8751	0.8976
Zhi's [23]	0.9087	0.9369	0.9245	0.9132	0.8943	0.9155
GALFusion [35]	0.8836	0.9154	0.9057	0.8880	0.8746	0.8935
SESF-Fuse [32]	0.8977	0.9332	0.9266	0.9011	0.8910	0.9099
MUFusion [34]	0.9139	0.9487	0.9401	0.9126	0.9055	0.9241
Proposed	0.9371	0.9769	0.9726	0.9381	0.9352	0.9520

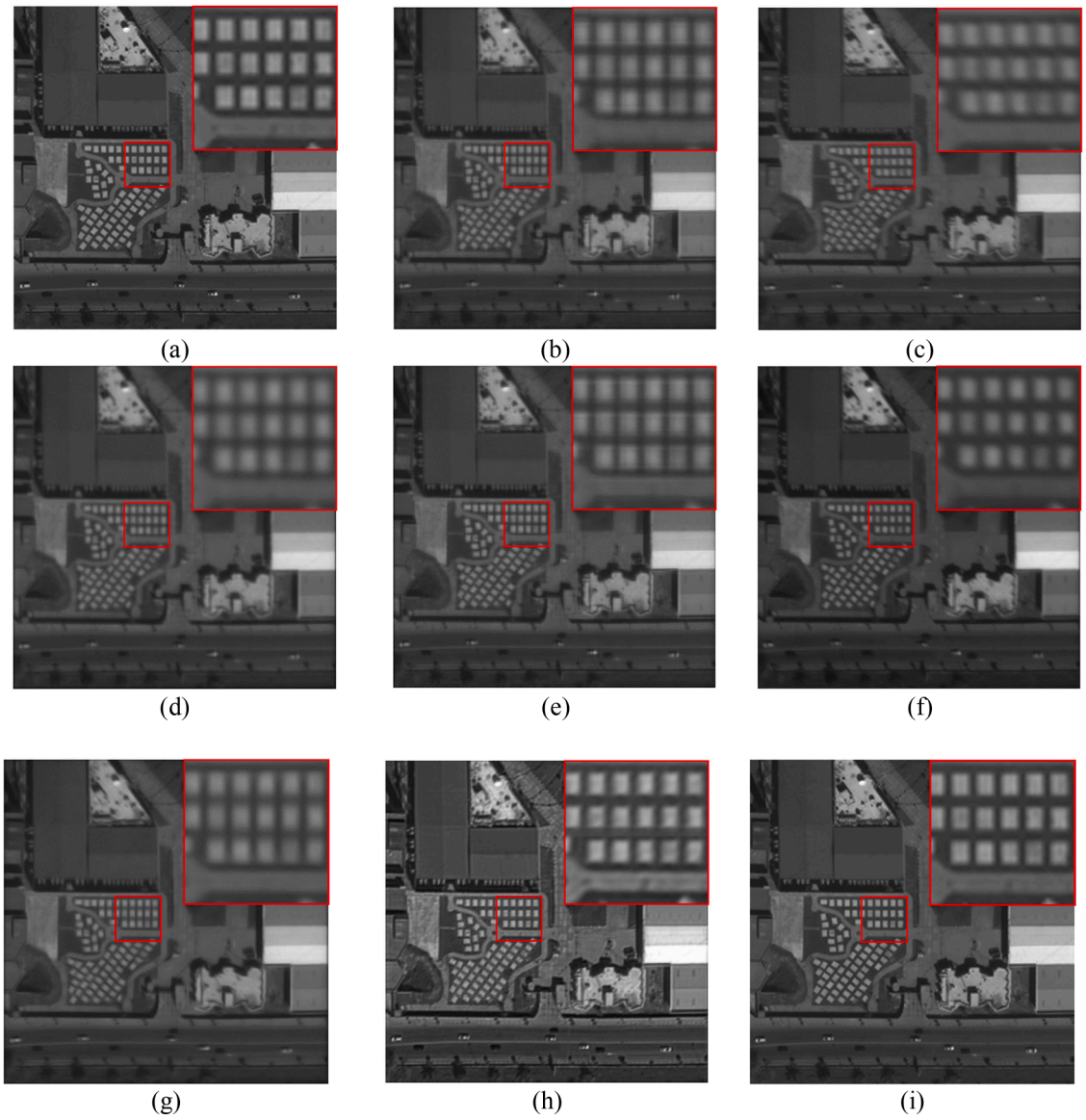

Fig. 13. Ground-truth and fusion results of the test image on residential area scene. (a) GT, (b)DWT, (c) PCA, (d) Zackay's, (e) Zhi's, (f) GALFusion, (g) SESF-Fuse, (h) MUFusion, and (i) proposed.

necessary to capture a minimum of N_{min} images evenly spaced at various angles to collect sufficient spatial frequency information in all directions. Specifically, N_{min} depends on the aspect ratio of the primary mirror, which can be determined using the following formula:

$$N_{\min} = \text{ceil}\left(\frac{180^{\circ}}{2 \times \arctan(b/a)}\right) = \text{ceil}\left(\frac{90^{\circ}}{\arctan(b/a)}\right) \tag{12}$$

where a and b are the length and width of the rectangle, respectively, and ceil$(\cdot)$ represents the ceiling function, which denotes the nearest integer greater than or equal to a given value.

To satisfy the lightweight requirements of the system while ensuring image restoration quality, [23] proposed that the aspect ratio of the rectangular aperture should be approximately three. Hence, we set b/a as $1/3$ in the digital simulation experiment by substituting this value into Eq. (12), we derive $N_{\min} = 5$. To introduce additional information into the time domain, we set the frame sequence length T to eight, with an angular interval of 22.5 °.

A semi-physical imaging experiment involves the utilisation of an imaging platform capable of simulating the RSA imaging process for imaging target scenes [44]. The captured images were then used for testing. The design scheme and physical diagrams are displayed in Figs. 9 and 10, respectively.

4.1.2. Implementation details

We set the window size to $8 \times 8 \times 8$ and the channel number C to 120. The head number of the multihead self-attention in RVAB is six, whereas in TMAB, it corresponds to a sequence length of eight. The Charbonnier loss function $\mathscr{L} = \sqrt{\| H - G \|^2 + \varepsilon^2}$ is used to calculate the loss, where G represents the ground-truth high-quality frame and H represents the fusion result, with ε set to 1×10^{-3}.

We compared our method with both general image-fusion methods and those specifically tailored to the RSA system. Among RSA-focused methods, we evaluated Zackay's approach, which is currently the only open-source method, and considered Zhi's method. Among the general image-fusion methods, our evaluation includes traditional techniques, such as discrete wavelet transformation (DWT) and principal component analysis (PCA), as well as state-of-the-art deep learning-based methods. These include multifocus image-fusion techniques, such as SESF-Fuse [32] and MUFusion [34], and multiexposure image-fusion techniques, such as GALFusion [35]. We used objective evaluation metrics such as

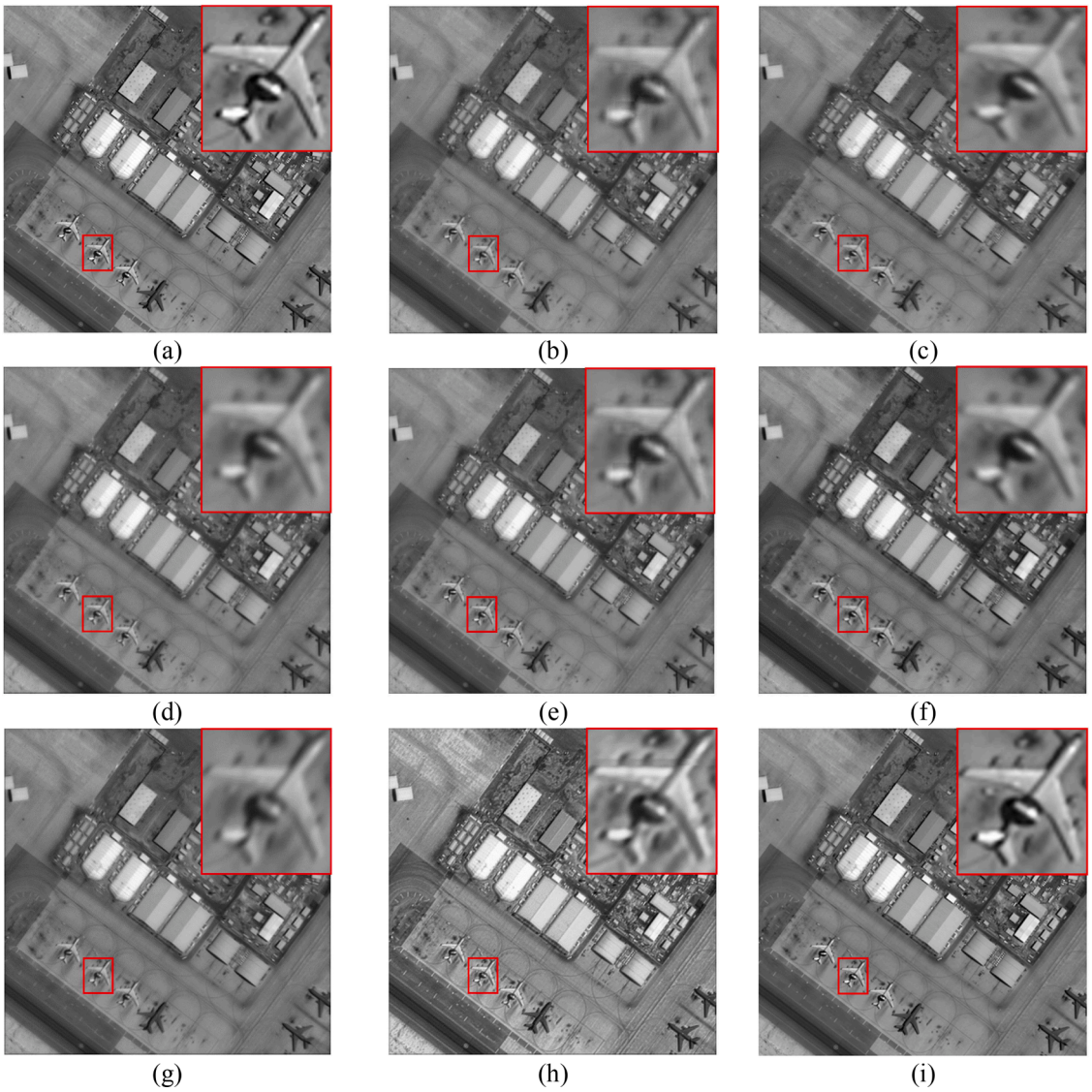

Fig. 14. Ground-truth and fusion results of the test image on airport scene. (a) GT, (b)DWT, (c) PCA, (d) Zackay's, (e) Zhi's, (f) GALFusion, (g) SESF-Fuse, (h) MUFusion, and (i) proposed.

structural similarity (SSIM) [45] and peak signal-to-noise ratio (PSNR), as well as subjective visual effects, to evaluate the fused output. The SSIM and PSNR are two metrics that assess the similarity between two images. SSIM considers the luminance, contrast, and structure of the images, providing a value between −1 and 1, where 1 indicates identical images, and lower values denote increasing dissimilarity. By contrast, PSNR is typically expressed in decibels (dB), and a higher PSNR value indicates greater similarity. Traditional methods, namely, DWT, PCA, and Zackay's methods, are registered based on the optical flow. The experimental process is shown in Fig. 11.

4.2. Experimental results and discussion

Tables 1 and 2 the results of the quantitative evaluation of the seven image-fusion methods mentioned above, as well as our proposed method. Fig. 12 provides a more intuitive representation of the average results, where the horizontal and vertical axes denote the SSIM and PSNR, respectively. As shown in the figure, the proposed method achieved the highest performance in terms of both the PSNR and SSIM metrics for the digital simulation test images containing five scenes. This demonstrates that our method can provide high-quality fused results with an accurate pixel intensity distribution and faithful texture and structure. Regarding the overall performance across all test images, our fusion results yielded a PSNR 32.77 dB and SSIM of 0.9520. This demonstrates an improvement of 1.25 dB in PSNR and 0.0279 dB in SSIM compared with MUFusion, which is the second-best approach. The proposed method fully considers the imaging mechanism and on-orbit characteristics of the RSA system, thereby effectively leveraging the powerful modelling capability of the transformer. Consequently, in scenes with high repetition rates of visual information, such as farmland and residential areas, the proposed method significantly outperformed the other methods. Notably, in the case of the farmland scene, our results achieve 35.70 dB and 0.9769 for PSNR and SSIM, respectively. This marks an impressive improvement of 6.16 % in the PSNR and 2.97 % in the SSIM compared with the second-best method.

We also present visual outcomes as a qualitative assessment of scenes, as shown in Fig. 4. Specifically, we show the locally magnified images and fusion results obtained using DWT, PCA, Zackay's method, Zhi's method, GALFusion, SESF-Fuse, MUFusion, and the proposed method in Figs. 13–15. The processing results of the semi-physical imaging experiment images are shown in Fig. 16. The image captured by the circular-aperture system and its locally magnified image are shown in Fig. 16(a). The fusion results of the DWT, PCA, Zackay's method, Zhi's method, GALFusion, SESF-Fuse, MUFusion, and the proposed method

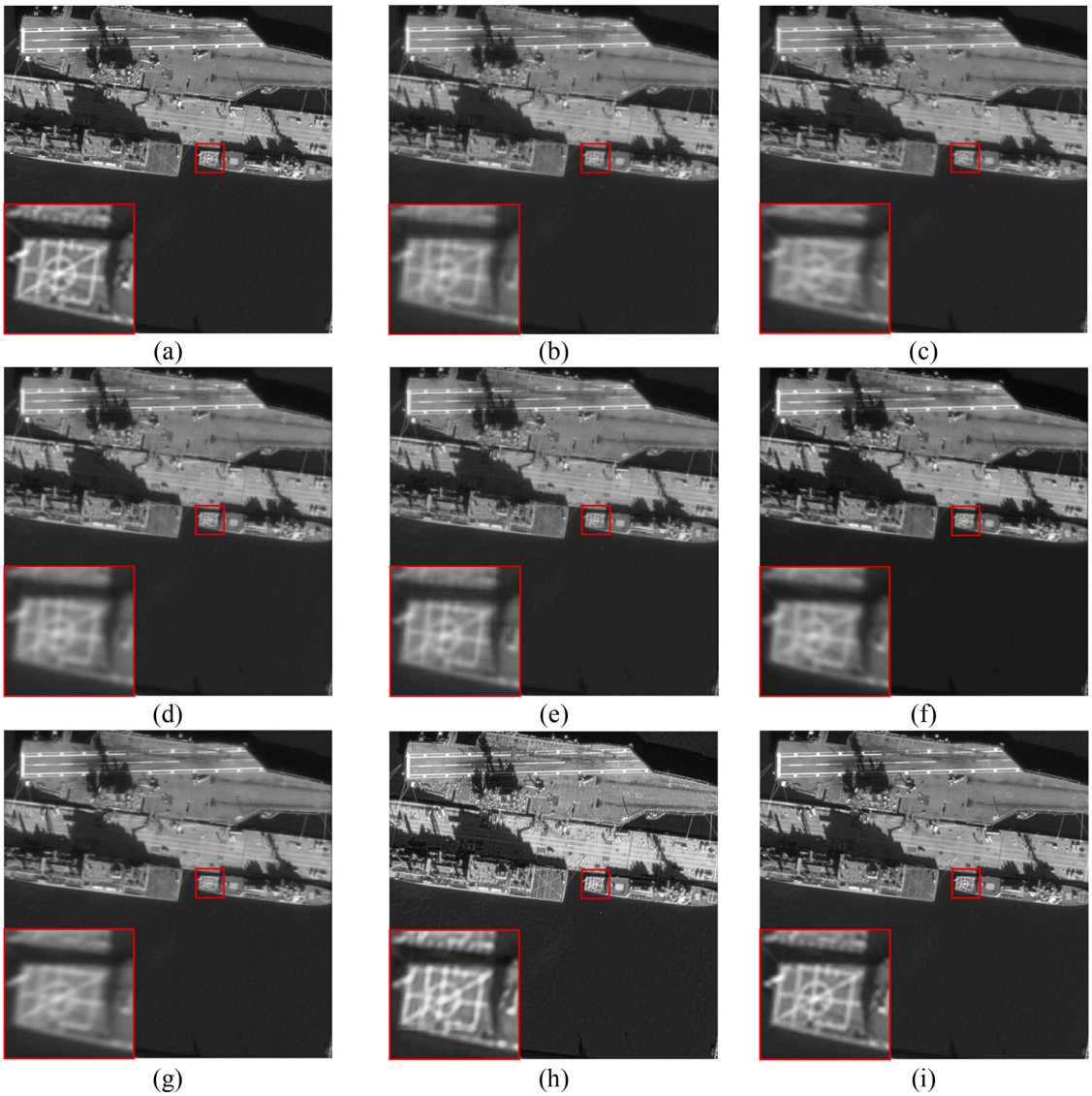

Fig. 15. Ground-truth and fusion results of the test image on harbour scene. (a) GT, (b)DWT, (c) PCA, (d) Zackay's, (e) Zhi's, (f) GALFusion, (g) SESF-Fuse, (h) MUFusion, and (i) proposed.

are illustrated in Fig. 16(b)–(i).

The above figures indicate that the fusion result of the PCA primarily mitigates the uneven resolution phenomenon without significantly enhancing the image clarity. Although DWT preserves image details better than PCA, it also retains artefacts introduced by explicit registration and motion compensation operations as well as high-frequency noise in the image sequence (e.g. the trailing effect observed in the square building edge in Fig. 13 and noise in the resolution target image in Fig. 16). Zackay's method limits the recovery of high-frequency information, resulting in a loss of detail and overly smoothed fused images with poor overall clarity. The fusion results of Zhi's method were significantly better than those of the three traditional methods mentioned above. However, it does not fully leverage the self-similarity of remote-sensing images, resulting in blurry textures in the fused images. For a fair comparison, we utilised publicly available codes for deep learning-based methods while maintaining the original architectures of the models designed for two-frame image input. However, because our input images consisted of eight frames, we integrated them sequentially in a straightforward manner. The visual results demonstrated that SESF-Fuse enhanced the clarity of small image patches within the images. Nevertheless, owing to its primary reliance on convolution, it exhibits limitations in terms of effectively utilising long-range dependencies. Furthermore, repeated fusion steps can lead to error accumulation, resulting in blurring of the final fused output. The collaborative aggregation module of GALFusion captures long-range pixel dependencies and mitigates artefacts. However, it can erroneously integrate the edges in the image, as illustrated in Fig. 13(f), leading to a slight reduction in the size of the fused buildings compared to their original size. MUFusion slightly surpasses Zhi's method in terms of objective evaluation metrics and notably improves subjective visual outcomes. Nevertheless, this method still struggles to effectively address the inter-frame displacement caused by satellite-platform vibrations. Consequently, the fusion results exhibited certain artefacts, as shown in Fig. 14(h) for the horizontal wing edges of the aircraft, and in Fig. 15(h) for the rear edges of the ship. By contrast, even without alignment modules, our fusion network based on ViT can adaptively establish connections between the most relevant pixels through attention, thereby effectively avoiding the generation of artefacts while obtaining more natural and reliable fusion results.

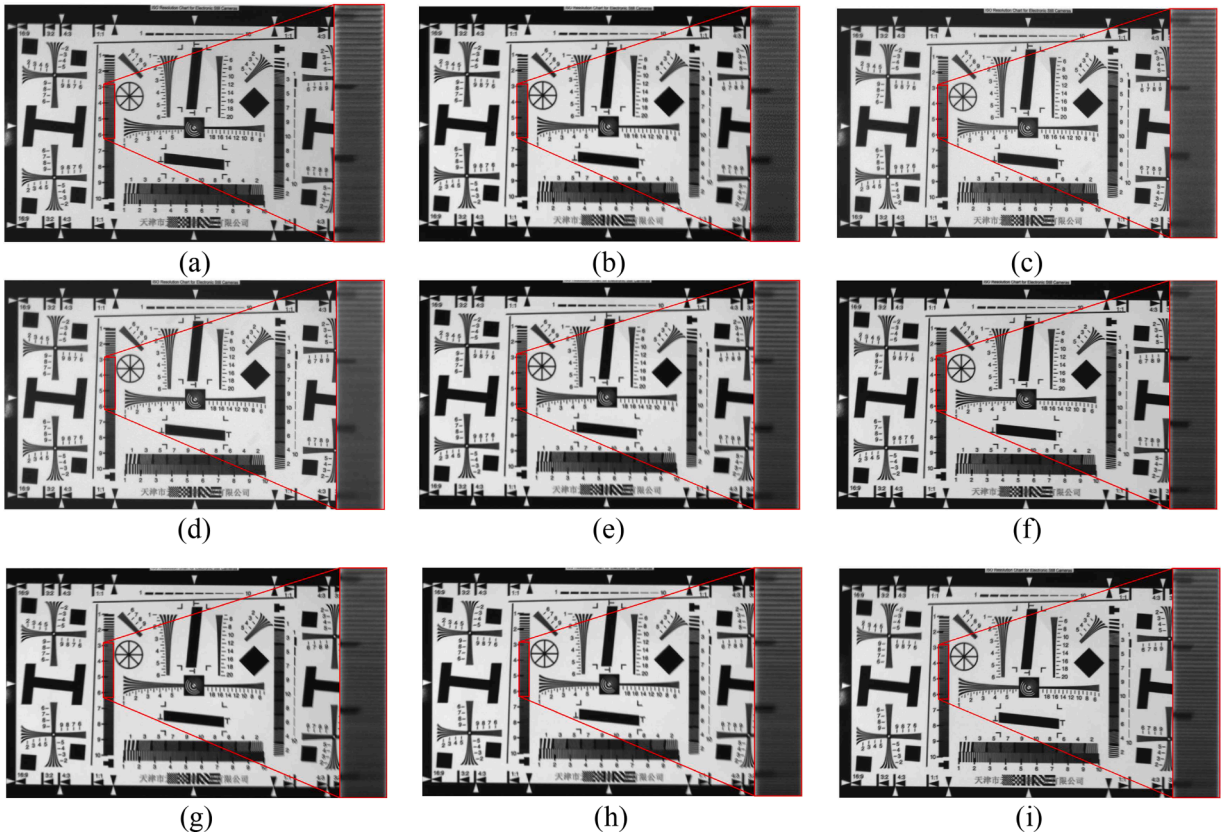

Fig. 16. Fusion results of the semi-physical experiment. (a) Circular primary mirror, (b)DWT, (c) PCA, (d) Zackay's, (e) Zhi's, (f) GALFusion, (g) SESF-Fuse, (h) MUFusion, and (i) proposed.

5. Conclusions

In this study, we propose an image-fusion method for the on-orbit imaging characteristics of an RSA system. First, we establish a mathematical model for a rectangular rotating pupil and analyse its on-orbit imaging characteristics. Subsequently, we propose an end-to-end image-fusion network based on ViT. The space–time domain feature-extraction module of the network comprises rotated attention blocks of various sizes and temporal mutual attention blocks. The RVAB generates windows with different locations, sizes, shapes, and angles based on window-based self-attention, which is beneficial for better processing the information of objects with different orientations and scales in remote-sensing images. TMAB uses inter-frame mutual attention instead of explicit alignment modules, which can adaptively capture the correlation between pixels in close proximity in different frames and reduce the generation of artefacts. Video swin transformer blocks are employed in the space–time domain feature fusion module of the network to fully fuse spatiotemporal information with the transformer's powerful modelling capability. The proposed method was compared with seven other methods, including the latest deep learning-based image-fusion method, using digital simulations and semi-physical imaging experiments with various scenes. The results demonstrate that the proposed method exhibits superior performance in both objective evaluation and image interpretation applications. Notably, our method achieved an overall performance of 32.77 dB in PSNR and 0.9520 in SSIM across all test images, showing a significant enhancement of 3.97 % in PSNR and 3.02 % in SSIM compared with MUFusion, the second-best approach. One limitation of the proposed method is its requirement for an adequate number of input frames to ensure the sampling of sufficient information in all directions, thereby supporting the reconstruction of high-quality fusion images. Therefore, in future research, our objective will be to explore single-image super-resolution methods for RSA systems to overcome this limitation.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC) [grant numbers 62305086, 62101160, 61975043] and the Innovation Foundation of CAST-BISEE [grant number CAST-BISEE2019–029].

CRediT authorship contribution statement

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

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.

References

[1] A.S. Belward, J.O. Skøien, Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites, ISPRS J. Photogramm. Remote Sens. 103 (2015) 115–128.

[2] X. Tong, J. Wang, G. Lai, J. Shang, C. Qiu, C. Liu, L. Ding, H. Li, S. Zhou, L. Yang, Normalized projection models for geostationary remote sensing satellite: a comprehensive comparative analysis (January 2019), IEEE Trans. Geosci. Remote Sens. 57 (12) (2019) 9643–9658.

[3] 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), Changchun, IEEE, 2018, pp. 1861–1866.
[4] X. Yang, F. Li, L. Xin, X. Lu, M. Lu, N. Zhang, An improved mapping with super-resolved multispectral images for geostationary satellites, Remote Sens. (Basel) 12 (2020) 466.
[5] Y. Wang, C. Zhang, L. Guo, S. Xu, G. Ju, Decoupled object-independent image features for fine phasing of segmented mirrors using deep learning, Remote Sens. (Basel) 14 (18) (2022) 4681.
[6] G. Chanan, D.G. MacMartin, J. Nelson, T. Mast, Control and alignment of segmented-mirror telescopes: matrices, modes, and error propagation, Appl. Opt. 43 (2004) 1223.
[7] S. Esposito, E. Pinna, A. Puglisi, A. Tozzi, P. Stefanini, Pyramid sensor for segmented mirror alignment, Opt. Lett. 30 (2005) 2572.
[8] S. Jiang, J. Hu, X. Zhi, W. Zhang, D. Wang, X. Sun, Local adaptive priorbased image restoration method for space diffraction imaging systems, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–10.
[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] S. Jiang, X. Zhi, Y. Dong, W. Zhang, D. Wang, Inversion restoration for space diffractive membrane imaging system, Opt. Lasers Eng. 125 (2020), 105863.
[11] D. Wang, X. Zhi, W. Zhang, Z. Yin, S. Jiang, R. Niu, Influence of ambient temperature on the modulation transfer function of an infrared membrane diffraction optical system, Appl. Opt. 57 (2018) 9096.
[12] 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.
[13] M.R. Rai, J. Rosen, Optical incoherent synthetic aperture imaging by superposition of phase-shifted optical transfer functions, Opt. Lett. 46 (2021) 1712.
[14] J. Wu, F. Yang, L. Cao, Resolution enhancement of long-range imaging with sparse apertures, Opt. Lasers Eng. 155 (2022), 107068.
[15] H. Touma, F. Martin, C. Aime, Image reconstruction using a rotating slit aperture telescope with partial atmospheric corrections, Pure Appl. Opt. 4 (1995) 685–694.
[16] 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.
[17] G. Nir, B. Zackay, E.O. Ofek, Can telescopes with elongated pupils achieve higher contrast and resolution?, in: Optical and Infrared Interferometry and Imaging VI, 10701 SPIE, 2018, pp. 200–204.
[18] B. Zackay, E.O. Ofek, A. Gal-Yam, Proper image subtraction—optimal transient detection, photometry, and hypothesis testing, ApJ 830 (2016) 27.
[19] B. Zackay, E.O. Ofek, How to COAAD images. i. optimal source detection and photometry of point sources using ensembles of images, Astrophys. J. 836 (2) (2017) 187.
[20] B. Zackay, E.O. Ofek, How to coaad images. ii. a coaddition image that is optimal for any purpose in the background-dominated noise limit, Astrophys. J. 836 (2) (2017) 188.
[21] H. Zhou, Y. Chen, H. Feng, G. Lv, Z. Xu, Q. Li, Rotated rectangular aperture imaging through multi-frame blind deconvolution with hyper-Laplacian priors, Opt. Express 29 (8) (2021) 12145–12159.
[22] G. Lv, H. Xu, H. Feng, Z. Xu, H. Zhou, Q. Li, Y. Chen, A full-aperture image synthesis method for the rotating rectangular aperture system using Fourier spectrum restoration, in: Photonics, 8, MDPI, 2021, p. 522.
[23] 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.
[24] J. Lin, Y. Cai, X. Hu, H. Wang, Y. Yan, X. Zou, H. Ding, Y. Zhang, R. Timofte, L. Van Gool, Flow-guided sparse transformer for video deblurring, arXiv preprint arXiv: 2201.01893 (2022).
[25] T. Gao, C. Wang, J. Zheng, G. Wu, X. Ning, X. Bai, J. Yang, J. Wang, A smoothing group lasso based interval type-2 fuzzy neural network for simultaneous feature selection and system identification, Knowl. Based Syst. 280 (2023), 111028.
[26] H. Zhang, S. Li, J. Qiu, Y. Tang, J. Wen, Z. Zha, B. Wen, Efficient and effective nonconvex low-rank subspace clustering via SVT-free operators, IEEE Trans. Circuits Syst. Video Technol. (2023), 1–1.
[27] 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), 1–1.
[28] P. Zhang, X. Yu, X. Bai, C. Wang, J. Zheng, X. Ning, Joint discriminative representation learning for end-to-end person search, Pattern Recognit. 147 (2024), 110053.
[29] Y. Sun, X. Zhi, H. Han, S. Jiang, T. Shi, J. Gong, W. Zhang, Enhancing UAV detection in surveillance camera videos through spatiotemporal information and optical flow, Sensors 23 (2023) 6037.
[30] 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 (2023) 6171.
[31] F. Xu, J. Liu, Y. Song, H. Sun, X. Wang, Multi-exposure image fusion techniques: a comprehensive review, Remote Sens. (Basel) 14 (2022) 771.
[32] B. Ma, Y. Zhu, X. Yin, X. Ban, H. Huang, M. Mukeshimana, SESF-Fuse: an unsupervised deep model for multi-focus image fusion, Neural Comput. Appl. 33 (2021) 5793–5804.
[33] K. Zheng, J. Huang, H. Yu, F. Zhao, Efficient Multi-exposure image fusion via filter-dominated fusion and gradient-driven unsupervised learning, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, IEEE, 2023, pp. 2805–2814.
[34] C. Cheng, T. Xu, X.-J. Wu, MUFusion: a general unsupervised image fusion network based on memory unit, Inf. Fusion 92 (2023) 80–92.
[35] J. Lei, J. Li, J. Liu, S. Zhou, Q. Zhang, N.K. Kasabov, GALFusion: multi-exposure image fusion via a global–local aggregation learning network, IEEE Trans. Instrum. Meas. 72 (2023) 1–15.
[36] J. Liu, G. Wu, J. Luan, Z. Jiang, R. Liu, X. Fan, HoLoCo: holistic and local contrastive learning network for multi-exposure image fusion, Inf. Fusion 95 (2023) 237–249.
[37] 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.
[38] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, L. Zhang, Advancing plain vision transformer towards remote sensing foundation model, IEEE Trans. Geosci. Remote Sens. 61 (2022) 1–15.
[39] S. Shi, J. Gu, L. Xie, X. Wang, Y. Yang, C. Dong, Rethinking alignment in video super-resolution transformers, arXiv preprint arXiv:2207.08494 (2022).
[40] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3202–3211.
[41] Y. Sun, X. Zhi, S. Jiang, J. Gong, T. Shi, N. Wang, Imaging simulation method for novel rotating synthetic aperture system based on conditional convolutional neural network, Remote Sens. (Basel) 15 (3) (2023) 688.
[42] 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.
[43] P. Yi, Z. Wang, K. Jiang, J. Jiang, J. Ma, Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3106–3115.
[44] 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 (2023) 17015.
[45] 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.