Recursive Self-Attention Modules-Based Network for Panchromatic and Multispectral Image Fusion

Chuang Liu, Lu Wei, Zhiqi Zhang, Xiaoxiao Feng, and Shao Xiang

***Abstract*—In the field of remote sensing, image fusion technology plays a crucial role in observing the state of global resources and environmental conditions, proposing response strategies, and constantly monitoring and correcting strategies. Currently, the majority of traditional methods exhibit varying degrees of spatial or spectral distortion, and these unreasonable spectral distributions may contain erroneous geographical feature information. Meanwhile, despite their performance in fusion results, deep learning-based methods cannot be applied to some practical application scenarios due to the requirement for hardware specifications resulting from a large number of parameters in their models. These issues are not conducive to accurately reflecting the actual geomorphic resource conditions or promoting sustainable development. In order to address the above issues, we propose a novel recursive self-attention module (RSAM), which consists of two stages: spatial-spectral similarity extraction and self-attention weight generation. The proposed RSAM employs a global-to-local strategy to capture the global interdependencies of two distinct local locations in the feature map. This method allows for simultaneous consideration of both spatial and spectral information while focusing on more mutual information between spectral and spatial dimensions. Subsequently, we construct the corresponding recursive self-attention residual block (RSARB) through RSAM and concatenate the RSARBs to generate a recursive self-attention module-based network (RSANet) with a limited number of parameters. Extensive experiments demonstrate that RSANet achieves superior results in both qualitative and quantitative evaluation despite the model parameters being within a narrow range of orders of magnitude. This demonstrates that the proposed method possesses robust feature learning capability and practicality for observing and studying the global resource environment.**

***Index Terms*—Earth observation, image fusion, remote sensing, self-attention.**

Manuscript received 21 August 2023; revised 5 October 2023; accepted 20 October 2023. Date of publication 24 October 2023; date of current version 14 November 2023. This work was supported in part by the National Key R&D Program of China Under Grant 2022YFB3902800, in part by the National Natural Science Foundation of China under Grant 61901307, and in part by the Scientific Research Foundation for Doctoral Program of Hubei University of Technology under Grant BSQD2020054 and Grant XJ2022005101. *(Chuang Liu and Lu Wei contributed equally to this work.) (Corresponding author: Zhiqi Zhang.)*

Chuang Liu and Xiaoxiao Feng are with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China (e-mail: liuchuang@hbut.edu.cn; 20220026@hbut.edu.cn).

Lu Wei is with the School of Information Science and Engineering, Wuchang Shouyi University, Wuhan 430064, China (e-mail: weilu@wsyu.edu.cn).

Zhiqi Zhang is with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zzq540@hbut.edu.cn).

Shao Xiang is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: xiangshao@whu.edu.cn).

The source code will be publicly available at https://github.com/JUSTM0VE0N/RSANet.

Digital Object Identifier 10.1109/JSTARS.2023.3327167

I. Introduction

OPTICAL remote sensing satellites are playing an increasingly prominent role as an essential means of acquiring spatial geographic information, which can rapidly acquire global ground images to realize the demand for Earth observation and can help us better understand and detect the Earth system, including the natural environment, ecosystems, and natural or human-induced disasters [1], [2], [3], [4]. However, the satellites cannot directly acquire high spatial resolution multispectral (HRMS) images because of some physical conditions. However, they are capable of acquiring paired high spatial resolution panchromatic (PAN) images and low spatial resolution multispectral (LRMS) images, respectively. Therefore, image fusion technology was derived, with the objective of combining two images of varying resolution to obtain HRMS images. The two main groups of fusion methods that can be classified are traditional and deep learning (DL) based.

Previously, most traditional methods relied on the manual extraction of PAN and LRMS features based on prior expertise. Nonetheless, due to the uncomplicated configuration of these features, their representational ability is limited, as evidenced by the fact that the results fused by these methods suffer from severe spatial or spectral distortions in some complex scenes. In fact, some classical traditional methods have been significantly improved, thereby preserving their fundamental characteristic of high computational efficiency [5], [6], [7], [8]. In recent years, deep neural networks have achieved incredible results in computer vision, and the novel paradigms in these networks have inspired and guided work in related fields [9], [10], [11], [12], [13]. Fig. 1 exhibits the relationship between the fusion performance and model parameters of our proposed method and the other five SOTA DL-based methods. In order to achieve better fusion results, more complex and in-depth architectures have been proposed. In addition, with the successful application of attention mechanisms in natural language processing, various enhanced transformer models and attention mechanisms have been applied to the field of remote sensing. However, they require a considerable amount of computational power to perform matrix multiplication operations, resulting in high model memory consumption and low efficiency.

It is widely acknowledged that the quality of the fused image is a crucial factor in determining the precision of subsequent tasks, such as classification and detection. However, unreasonable spatial and spectral distributions may hinder the subsequent tasks;

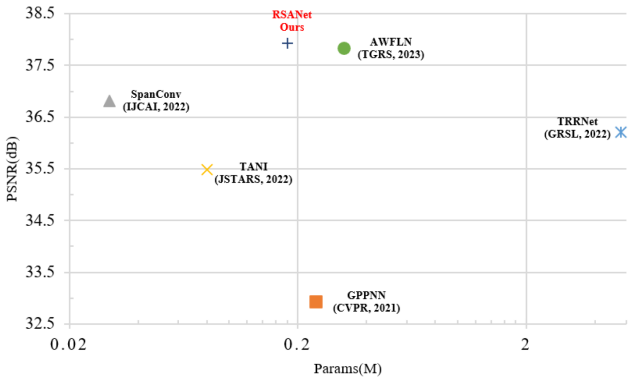

Fig. 1. PSNR-Params comparisons between five SOTA DL-based fusion methods and our RSANet on the QuickBird satellite dataset. The vertical axis is PSNR (fusion performance), and the horizontal axis is Params (model complexity).

hence, it is imperative to ensure the fusion effect of the image. Furthermore, in practical scenarios, having fewer parameters can reduce the device specification requirement for the fusion method. This, in turn, provides a more flexible margin within the same specifications for subsequent tasks [14]. On the other hand, it is capable of dividing each computational unit within the embedded device into larger units, leading to enhanced performance in constrained application scenarios, such as on-board processing [15], [16], [17], [18], [19]. Therefore, we are committed to achieving good fusion results while ensuring that the model parameters are within a modest order of magnitude.

A. Our Motivation

Recently, the application of DL algorithms to image fusion has received widespread attention. Several instances of data processing results in publicly available literature demonstrate that DL algorithms can substantially enhance the performance of satellite image fusion processing. Nevertheless, to enable on-board satellite image fusion using DL, it is imperative to reduce the model complexity. Existing fusion algorithms have high computational complexity, making it challenging to apply them to computation-constrained platforms. There are also lightweight methods with low computational complexity, but their fusion accuracy is not sufficient for subsequent tasks, such as classification [20]. Attention mechanisms can effectively enhance the fusion capability of the model and have a significant advantage in capturing long-range dependencies. This motivated us to employ it to establish long-distance correlations between the spectral information in LRMS images and the spatial information in PAN images. Furthermore, we reduce the computational complexity of the attention module to ensure the minimum number of parameters in the entire fusion model without compromising fusion accuracy. This enables our method to be applicable in on-board environments.

B. Differences From Other Related Works

To the best of our knowledge, attention mechanisms have been applied to a variety of visual tasks, including image compression [21], detection [22], [23], and image classification [24]. In the domain of image fusion, there are two issues with these methods: 1) The majority of the existing attention modules focus only on the information between spectral bands or information between spatial dimensions. Subsequently, different modules are incorporated into a specific network structure, which ignores the mutual information across both spectral and spatial dimensions. 2) The majority of the modeling structures are constructed to be extremely intricate, which results in a model with greater initial complexity. Hence, it is imperative to devise a straightforward and efficacious fusion model utilizing the attention mechanism. In this regard, we propose the recursive self-attention module (RSAM) to integrate the respective information of PAN and LRMS images as well as the correlation information between them. Based on this foundation, instead of devising a more intricate network architecture, we employ a straightforward and efficient method for embedding RSAM, limiting the parameters of the model to a smaller order of magnitude.

To achieve an optimal balance between the performance of the fusion and the number of model parameters, we develop a novel RSAM, which consists of two stages: spatial-spectral similarity extraction and self-attention weight generation. Furthermore, we construct a straightforward single-branch network to comprehensively assess the efficacy of the proposed RSAM, commonly referred to as a recursive self-attention module-based network (RSANet). The main contributions of this work can be summarized as follows.

1) In contrast to existing methods that focus on either the spectral bands or spatial dimensions, this paper introduces a novel RSAM for image fusion tasks. It adopts a global-to-local (G2L) strategy to integrate the feature representations among each spectral band and spatial pixel.
2) The customized G2L strategy establishes global dependencies by calculating the similarity between individual spectral bands and between spatial pixel points. Subsequently, local representations are established based on the global information to capture the features at different resolutions, thereby enhancing the recovery of spatial details in LRMS images.
3) We propose a straightforward end-to-end network, RSANet, based on the concept of super-resolution. The effectiveness of our method is validated by evaluating it on three available datasets: QB, GF-2, and WV-3. The experimental results demonstrate that our method achieves state-of-the-art fusion performance while utilizing fewer parameters. In comparison to traditional methods, our method effectively preserves both spatial details and spectral information. Moreover, when compared with DL-based methods, our method demonstrates a better balance between the fusion performance and the model parameters.

II. Related Work

A. Traditional Image Fusion Methods

The traditional methods for PAN and LRMS image fusion can be categorized into three distinct groups: component substitution (CS) based methods, multi-resolution analysis (MRA) based methods, and variational optimization (VO) based methods. The main goal of CS-based image fusion methods is to replace the low spatial resolution component (LR) in LRMS with PAN, whereas LR is generally obtained through the transformation of

LRMS images. Many pioneering fusion methods are based on the CS, such as intensity-hue-saturation [25], Brovey transformation [26], principal component analysis, and Gram–Schmidt (GS). As the comprehension of the task of fusion grew, numerous new methods departed from the complexities of rigid projection transformations, and many enhanced variants of this concept were derived, including the Brovey transform with haze correction [27], the adaptive GS (GSA) [28], band-dependent spatial detail (BDSD) [29], and partial-replacement adaptive CS (PRACS) methods [30]. The CS-based methods usually yield superior visual effects, but they also introduce some spectral distortion problems. The MRA-based image fusion method involves performing spatial decomposition of PAN through some transformation to extract the high spatial structure and incorporate it into the interpolated LRMS to obtain HRMS. Commonly known MRA-based methods include additive wavelet luminance proportional (AWLP) [31], GLP with modulation transfer function (MTF) matched filter (MTF-GLP) [32], and smoothing filter-based intensity modulation [33]. The MRA-based method surpasses the CS-based method in terms of fusion results. However, it presents the challenge of spatial distortion. The VO-based method is based on variational theory, which considers the image fusion task to be an ill-posed problem and applies a specified prior term to regulate the potential HRMS [34], [35]. This method has good performance in spatial and spectral preservation. Nonetheless, the limitations of these methods lie in their capability of tuning numerous hyperparameters and shallow nonlinear representations [36], [37].

B. Deep Learning-Based Methods

The DL-based methods, in contrast to the three previous traditional methods, are capable of extracting deep semantic information from PAN and LRMS images, which greatly aids the subsequent feature fusion. Meanwhile, the process of remote sensing technology has enabled the accumulation of numerous global surface images to be utilized for the training of neural networks, giving the model a nonlinear fitting capability that is not present in traditional methods. Specifically, Masi et al. [38] utilized three convolution layers to learn the mapping of PAN and LRMS images to HRMS. Despite achieving satisfactory fusion results at that time, the three-layer network structure lacked sufficient fitting capability of fully utilizing the spatial information present in the PAN images. After savoring the savor of neural networks, numerous fusion methods based on DL were proposed. Among them, Yang et al. [39] designed PanNet using residual structure. The successful amalgamation of multiple residual blocks with the notion of detail injection results in a fusion result that exhibits extensive spatial details. Nevertheless, this method requires a substantial number of additional parameters. Xu et al. [40] devised a model-driven fusion network influenced by the VO problem, GPPNN, which achieves excellent fusion results while maintaining a certain range of parameters. In addition, Chen et al. [41] devised a lightweight network by exploring the kernel space.

C. Attention Mechanism

In general, the attention mechanism prioritizes the important information among the numerous features and disregards the majority of the irrelevant information. Attention mechanisms are now extensively used in computer vision. Hu et al. [42] proposed a channel attention mechanism that utilized a squeeze excitation (SE) module to capture interchannel relationships. In contrast to the SE channel, which solely focuses on channel information, Woo et al. [43] combined spatial and channel attention to propose a lightweight and generalized attention mechanism module that represents a convolutional block (CBAM). Liu et al. [44] proposed a temporal adaptive module, considering the importance of cross-dimensional interactions. Liu et al. [45] redesigned a global attention mechanism (GAM) using CBAM. Furthermore, in the field of image fusion, Diao et al. [46] utilized the Information Interaction Module (IIM) to exchange the feature maps in the spatial attention subnetwork and the spectral attention subnetwork, thereby ensuring compatibility between the feature maps. In order to reduce the redundancy between features, Diao et al. [47] proposed a fusion method combining convolutional and Swin-transformer blocks. Lu et al. [48] utilized an adaptive feature learning block based on spatial-spectral interleaved attention to construct a lightweight fusion network.

To summarize, the majority of DL-based methods can achieve superior fusion results in comparison to CS-, MRA-, and VO-based methods. However, these methods necessitate additional storage space and hardware device support.

III. Methodology

In this section, we begin by defining the mathematical notations used in this article. Subsequently, the procedure for fabricating the proposed RSAM is explained. Following this, we construct the basic recursive self-attention residual block (RSARB) through RSAM. Finally, we embed these RSARBs into RSANet, which is capable of obtaining details in different dimensions of images while preserving more spatial details and spectral information during the fusion process.

A. Notations

Let PAN and LRMS images be represented by $P \in R^{1\times rh\times rw}$ and $M \in R^{b\times h\times w}$, where b represents the number of bands in LRMS images, h and w represent the height and width of the image, respectively. r is the ratio of spatial resolution between PAN and LRMS images. In our proposed network, we up-sample M by a factor of 4 to obtain $\bar{M} \in R^{b\times rh\times rw}$, and $\tilde{M} \in R^{b\times rh\times rw}$ is obtained by fusing P with M. Furthermore, in the reduced-resolution (RR) testing, the ground truth (GT) is denoted as $G \in R^{b\times rh\times rw}$.

B. RSAM

As remote sensing images may exhibit significant variances between homogeneous areas and minor variances between heterogeneous areas, it is imperative that the traditional convolutional network incorporate multiple layers to expand the local

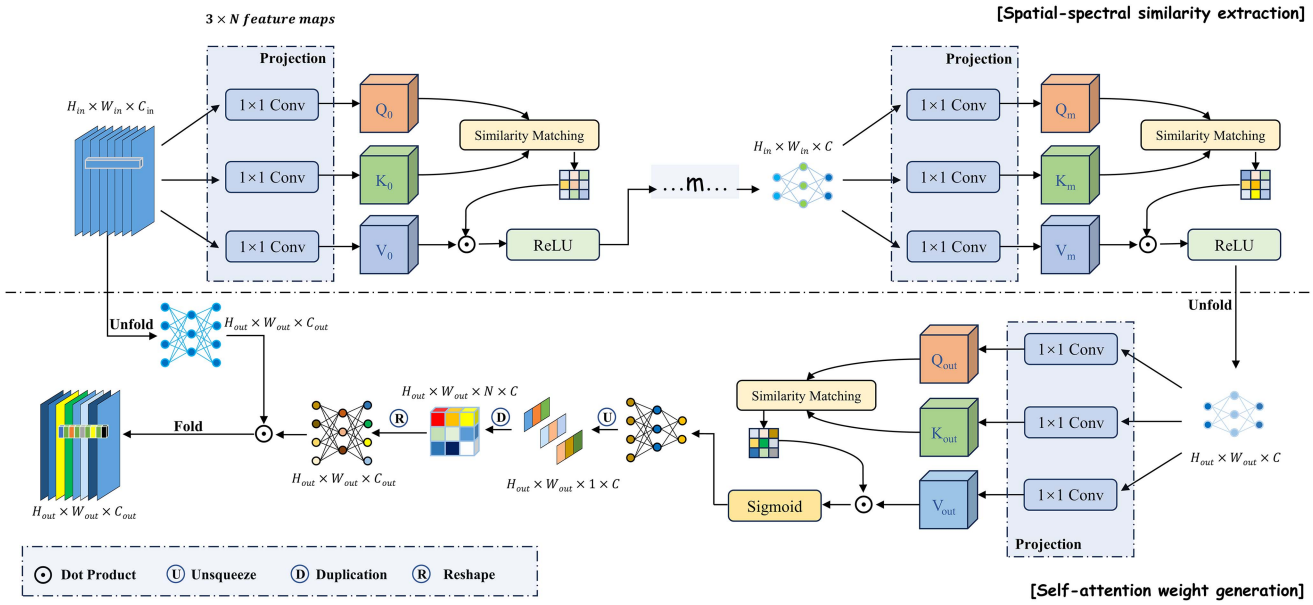

Fig. 2. Overview of the RSAM architecture.

receptive field to the global area. This necessitates the accumulation of multiple layers when capturing long-range feature dependencies, resulting in a low level of learning efficiency. Furthermore, transformer-based and attention-based methods use the attention mechanism to establish long-range dependencies and focus on important regions within a larger feature, but there are too many matrix operations in their models, which leads to a large amount of feature redundancy.

We propose the RSAM, which consists of two stages: spatial-spectral similarity extraction and self-attention weight generation. We employ a G2L strategy to integrate the feature representations between each spectral band and between each spatial pixel point. In the first stage, we calculate the values between different spectral bands and between each spatial pixel point through a finite number of recursive attention mechanisms in order to strengthen the connection between image regions with high similarity. This will establish the global relationship between LRMS and PAN images. In the interim, we incorporate the ReLU activation function into each attention module to enhance the nonlinear mapping capability of the model while simultaneously eliminating insignificant information. As shown in Fig. 2, in the spatial-spectral similarity extraction stage, we project the input features into queries, keys, and values by using three 1×1 convolutions. This can be formulated as

$$Q_{ij}^{(1)}, K_{ij}^{(1)}, \ V_{ij}^{(1)} = \text{Conv}\left(F_{ij}^{\text{in}}\right), \text{Conv}\left(F_{ij}^{\text{in}}\right), \text{Conv}\left(F_{ij}^{\text{in}}\right) \tag{1}$$

where F_{ij}^{in} denotes the corresponding tensor of the input feature at pixel (i, j). Then the similarity weights after similarity matching between the query and each key, and finally obtain the similarity between different spectral channels and spatial pixel points by performing dot product operations and nonlinear mapping on the generated weights and values. Specifically, it can be formulated as

$$A\left(Q_{ij}^{(1)}, K_{ab}^{(1)}\right) = \text{Softmax}\left(\left(Q_{ij}^{(1)}, K_{ab}^{(1)}\right) + \text{Conv}\left(P_1\right)\right) \tag{2}$$

$$F_{ij}^{(1)} = \text{ReLU}\left(\sum_{a,b\in R(i,j)} A\left(Q_{ij}^{(1)}, K_{ab}^{(1)}\right) V_{ab}^{(1)}\right) \tag{3}$$

where P_1 represents a learnable parameter to represent the position bias, and $R(i, j)$ represents the local pixel region centered at (i, j). Considering that remote sensing images may exhibit significant variances between homogeneous areas and minor variances between heterogeneous areas, we believe that relying on only one self-attention module is not able to extract the accurate similarity between each spectral band and between each spatial pixel point, so we stack m self-attention modules for extracting the accurate similarity between deep semantic features. The similarity extraction process for the mth time can be expressed as

$$Q_{ij}^{(m)}, K_{ij}^{(m)}, \ V_{ij}^{(m)} = \text{Conv}\left(F_{ij}^{m-1}\right), \text{Conv}\left(F_{ij}^{m-1}\right), \text{Conv}\left(F_{ij}^{m-1}\right) \tag{4}$$

$$A\left(Q_{ij}^{(m)}, K_{ab}^{(m)}\right) = \text{Softmax}\left(\left(Q_{ij}^{(m)}, K_{ab}^{(m)}\right) + \text{Conv}\left(P_m\right)\right) \tag{5}$$

$$F_{ij}^{(m)} = \text{ReLU}\left(\sum_{a,b\in R(i,j)} A\left(Q_{ij}^{(m)}, K_{ab}^{(m)}\right) V_{ab}^{(m)}\right). \tag{6}$$

The second stage involves unfolding the global similarity features obtained in the first stage and then calculating the similarity

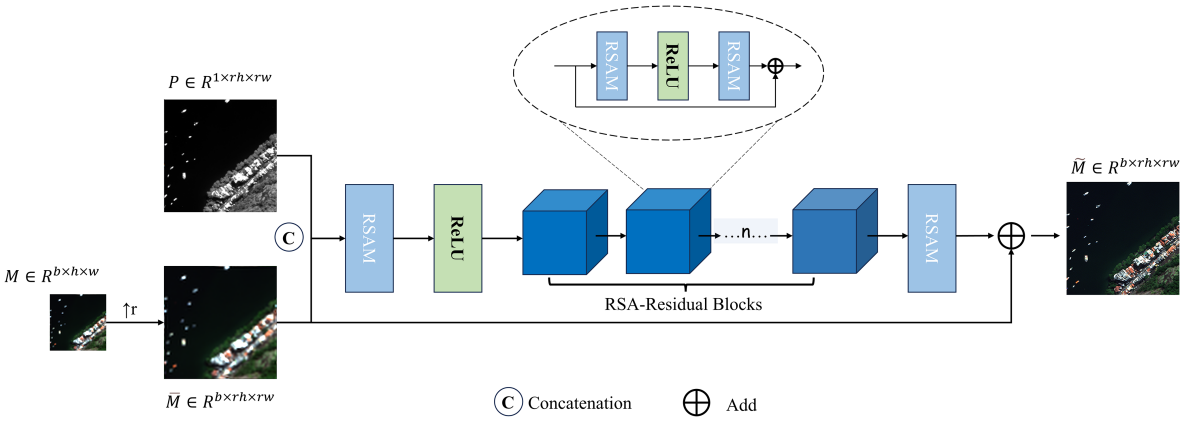

Fig. 3. Overall architecture of the proposed RSANet for image fusion.

of the local features. We then accurately represent the different importance of these relatively important features in the current localization through the sigmoid activation function. Finally, we obtain the self-attention weights for each localization. This can be formulated as

$$Q_{ij}^{(o)}, K_{ij}^{(o)}, \ V_{ij}^{(o)} = \text{Conv}\left(F_{ij}^{m'}\right), \text{Conv}\left(F_{ij}^{m'}\right), \text{Conv}\left(F_{ij}^{m'}\right) \tag{7}$$

$$A\left(Q_{ij}^{(o)}, K_{ab}^{(o)}\right) = \text{Softmax}\left(\left(Q_{ij}^{(o)}, K_{ab}^{(o)}\right) + \text{Conv}\left(P_o\right)\right) \tag{8}$$

$$W_{ij}^{o} = \text{Sigmoid}\left(\sum_{a,b \in R(i,j)} A\left(Q_{ij}^{(o)}, K_{ab}^{(o)}\right) V_{ab}^{(o)}\right) \tag{9}$$

$$F_{ij}^{o} = W_{ij}^{o} \odot F_{ij}^{\text{in}'} \tag{10}$$

where $F_{ij}^{\text{in}'}$ and $F_{ij}^{m'}$ are the unfolded forms of the input features and the output features of the first stage, respectively. W_{ij}^{o} is the corresponding attention weight at (i, j). F_{ij}^{o} denotes the corresponding tensor of the output feature at pixel (i, j).

C. RSARB

As shown in the dashed ellipse box in Fig. 3, we construct an RSARB on the basis of the proposed RSAM. The purpose of constructing RSARB is that there are two aspects, one of which is the incorporation of the ReLU activation function between two RSAM modules, which will allow RSAM to be more focused on extracting useful information in the input features. The other is the adoption of the classical residual structure, which will prevent the gradient from disappearing while preserving the crucial information extracted from the previous module. We avoid designing a more complex residual structure, instead employing the same structure as the original ResBlock [49] to fully evaluate the effectiveness of RSAM. We replace the convolutional layers in the original ResBlock with our proposed RSAM to form RSARB.

D. RSANet

We designed an RSANet, which is a single-branch network that is based on the concept of super-resolution [50]. We combine the PAN and the LRMS after up-sampling by a factor of 4 as the input to the network. Initially, the important information of the paired images is extracted by a single RSAM + ReLU, followed by the feeding of n RSAMs to perform adequate feature extraction and fusion (n is a hyperparameter, and we set n to be 8 in order to balance the effect of the fusion and the parameters of the network). Finally, the important information is reconstructed by a single RSAM. In this way, RSANet can be represented as

$$f = \overbrace{\text{RSAM}}^{\times n}\left(\left[P, \bar{M}\right]\right). \tag{11}$$

We consider RSANet as an underfitted nonlinear function f, so the ultimate goal of this function is to produce HRMS images $\tilde{M}$. Moreover, a global residual was designed based on f in order to preserve more spectral information. The final $\tilde{M}$ was obtained by summing $\bar{M}$ (up-sampled multispectral) with the output of the network, thus this process can be represented as

$$\tilde{M} = f_\theta\left(P; \bar{M}\right) + \bar{M} \tag{12}$$

where f_θ denotes the fitted network. In supervised learning, the fitted network is generally obtained by minimizing the supervised loss. As mentioned before, in order to fully verify the potential of RSAM, we do not design a more complex network structure based on RSAM to achieve better fusion results. Similarly, we simply follow the mean square error to constrain the network training

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} ||f_\theta\left(P_i; \bar{M}_i\right) + \bar{M}_i - G_i||_2 \tag{13}$$

where N represents the number of samples in the training set. G represents the ground truth, which is the goal of model learning. Therefore, we use the neural network to fit the nonlinear function to obtain our image fusion model.

TABLE I
DETAILS INFORMATION OF THE QUICKBIRD, GAOFEN-2, AND WORLDVIEW-3 DATASETS

	Bit depth	Resolution		Size						Number		
		P	*M*	P_d	M_d	P_{rr}	M_{rr}	P_{fr}	M_{fr}	*Train*	*Val*	*Test**
QB	11	0.61	2.44	64	16	256	64	512	128	17139	1905	40
GF-2	10	0.8	3.2	64	16	256	64	512	128	19809	2201	40
WV-3	11	0.3	1.2	64	16	256	64	512	128	9714	1080	40

QB, GF-2, and WV-3 are abbreviations for QuickBird, GaoFen-2, and WorldView-3, respectively. P and M Represent PAN and LRMS, respectively. P_d and M_d, P_{rr} and M_{rr}, and P_{fr} and M_{fr} represent the size of the PAN and LRMS image pairs at training, reduced resolution testing, and full resolution testing, respectively. * In the training set and validation set, the sizes of *M* and *P* are 16×16 and 64×64, respectively. In addition, the test set is divided into the RR test set and the FR test set, and each part contains 20 sets of image pairs. The corresponding sizes in the RR test set are 64×64 and 256×256, and the corresponding sizes in the FR test set are 128×128 and 512×512, respectively.

IV. COMPARATIVE EXPERIMENT

Extensive experiments are designed to examine the efficacy of our proposed RSAM. First, we furnish comprehensive experimental settings that encompass datasets, evaluation metrics, comparison methods, and implementation details. Subsequently, we evaluate our methodology on three publicly available satellite datasets and compare it with state-of-the-art methods.

A. Datasets

In order to contribute to sustainable development to some extent, images from three satellites were selected as datasets to thoroughly evaluate the performance of the RSAM, namely QuickBird (QB), GaoFen-2 (GF-2), and WorldView-3 (WV-3) [51]. All of them contain rich information on ground objects, such as urban roads, geographic buildings, and parked vehicles. Among them, LRMS images taken by QB and GF-2 satellites are both in four bands, and LRMS images taken by WV-3 are in eight bands. In all datasets, the spatial resolutions of the corresponding PAN images are 0.61, 0.8, and 0.3 m, while the spatial resolutions of the corresponding LRMS images are 2.44, 3.2, and 1.2 m. The details of the three datasets are given in Table I. Since the actual HRMS images were not available, following Wald's protocol [52], the images are first filtered by using the MTF of each satellite, followed by downsampling with the nearest interpolation to obtain the PAN and LRMS image blocks on the low-resolution scale, and the original LRMS images are used as the GT to ultimately obtain the datasets used for the training and validation of the network. In particular, the simulation of the test data is also the same as this process, and the test sets are divided into a simulated dataset and a real dataset, which are used for RR and full-resolution (FR) testing, respectively.

B. Evaluation Metrics

As previously stated, in accordance with Wald's protocol, we perform downsampling of both the original PAN and the LRMS images and utilize the original LRMS images as the GT, so five widely used reference metrics were selected to measure the performance of the method, which are structural similarity index measure (SSIM) [53], spectral angle mapper (SAM) [54], correlation coefficient (CC) [55], Erreur Relative Global Adimensionnelle de Synthese (ERGAS) [56], and peak signal-to-noise ratio (PSNR) [57]. Moreover, we test the proposed RSAM directly on the original PAN images and LRMS images, i.e., FR testing, so we choose five nonreference reference metrics, which are spectral distortion index (D_λ and D_λ^F), spatial distortion index (D_s), and nonreference quality assessment [(QNR) and hybrid quality with no reference (HQNR)] [58]. Since the QNR protocol may consider details injected during the fusion process to be spectral distortions, we employ D_λ and D_λ^F to comprehensively evaluate the spectral quality of the fused image. In turn, QNR and HQNR are employed to comprehensively evaluate the overall quality of the fused image [59]. Following are specific descriptions of ten metrics.

1) Structural similarity index measure [53]

$$\begin{aligned}&\mathrm{SSIM}\left(x_1, x_2\right)\\ &\triangleq \frac{\left(2\mu_{x_1}\mu_{x_2}+c_1\right)\left(2\sigma_{x_1}\sigma_{x_2}+c_2\right)\left(\sigma_{x_1x_2}+c_3\right)}{\left(\mu_{x_1}^2+\mu_{x_2}^2+c_1\right)\left(\sigma_{x_1}^2+\sigma_{x_2}^2+c_2\right)\left(\sigma_{x_1}\sigma_{x_1}+c_3\right)}\end{aligned} \tag{14}$$

where x_1 and x_2 denote the fused image and GT, μ and σ denote the mean and standard deviation of the image, respectively, and c_1, c_2, and c_3 denote unequal constants. SSIM is mainly used to measure the loss of information and distortion between x_1 and x_2, with an ideal value of 1, which means that the value is approximately close to 1, the better the information preservation of the fused image.

2) Spectral angle mapper [54]

$$\mathrm{SAM}\left(x_1, x_2\right) \triangleq \arccos\left(\frac{x_1 \cdot x_2}{\|x_1\| \cdot \|x_2\|}\right) \tag{15}$$

where arccos is the inverse cosine function, which is used to measure the angle between the spectral vectors of x_1 and x_2 under the corresponding pixel, so the SAM is mainly used as an assessment of the spectral distortion of the fused results, with an ideal value of 0, i.e., the closer the value is to 0, the better the spectral preservation of the fused image is.

3) Correlation coefficient [55]

$$\mathrm{CC}\left(x_1, x_2\right) \triangleq \frac{\mathrm{cov}\left(x_1,\ x_2\right)}{\mu_{x_1} \cdot \mu_{x_2}} \tag{16}$$

where cov is the covariance. The CC is mainly used to measure the correlation between x_1 and x_2. The ideal value of CC is 1.

4) Erreur Relative Global Adimensionnelle de Synthese [56]

$$\mathrm{ERGAS} \triangleq 100 \cdot r \sqrt{\frac{1}{b}\sum_{i=1}^{b}\frac{\mathrm{RMSE}^2\left(b_i\right)}{M^2\left(b_i\right)}} \tag{17}$$

where r denotes the ratio of the spatial resolution of the PAN images to that of the LRMS images, RMSE (b_i) is the root-mean-square error between the ith band of x_1 and x_2. $M(b_i)$ is the mean value of the original LRMS band b_i. ERGAS is mainly used to measure the error and the dynamic range situation between x_1 and x_2, and its ideal value is 0.

5) Peak signal-to-noise ratio [57]

$$\mathrm{PSNR}(x_1, x_2) \triangleq 10\log_{10}\frac{p_{\max}^2}{(x_1 - x_2)^2} \tag{18}$$

where $p_{\max}^2$ denotes the square of the largest possible pixel value in an image. PSNR is mainly used to measure the distortion of an image at the pixel level. The larger its value, the better.

6) Spectral distortion index (D_λ)

$$D_\lambda = \sqrt[p]{\frac{1}{b(b-1)}\sum_{i=1}^{b}\sum_{j=1}^{b}\left|Q(x_i, x_j) - Q(M_i, M_j)\right|^p} \tag{19}$$

where p is a positive integer, Q is the Q index, and M is LRMS.

7) Spectral distortion index (D_λ^F)

$$D_\lambda^F = 1 - Q2^n(x_{1\downarrow}, M) \tag{20}$$

where $Q2^n$ is a multivariate version of the Q index, and $x_{1\downarrow}$ is the result of the spatial degradation of the fused image x_1 by the MTF matched filter. The spectral distortion between x_1 and LRMS is measured using D_λ and D_λ^F in the absence of a GT. The smaller its value, the better.

8) Spatial distortion index (D_s)

$$D_s = \sqrt[q]{\frac{1}{b}\sum_{i=1}^{b}\left|Q(x_i, P) - Q(\bar{M}_i, P)\right|^q} \tag{21}$$

where q is a positive integer and P is PAN. The loss of spatial detail is measured using D_s in the absence of a GT. The smaller its value, the better.

9) Quality with no reference

$$\mathrm{QNR} = (1 - D_\lambda)^\alpha (1 - D_s)^\beta \tag{22}$$

QNR is obtained by weighting D_λ and D_s.

10) Hybrid quality with no reference [58]

$$\mathrm{HQNR} = \left(1 - D_\lambda^F\right)^\alpha (1 - D_s)^\beta. \tag{23}$$

The overall quality of fused images is measured by QNR and HQNR, and its ideal value is 1.

C. Comparative Methods

In our comparison experiments, 14 state-of-the-art methods are selected to compare with our proposed method. These methods comprise three CS-based methods, namely GSA [28]. PRACS [30], and BDSD-PC [60]. In addition, we select three MRA-based methods, namely AWLP [41], MTF-GLP-FS [61], and MTF-GLP-HPM-R. Furthermore, we choose eight DL methods, including PNN [38], PanNet [39], BDPN [62], GPPNN [40], SpanConv [41], TANI [46], TRRNet [47], and AWFLN [48].

D. Training Details

All the comparative methods are trained and tested on the same device to ensure fairness. We use Python3.7 and the PyTorch framework to implement the methods on the Ubuntu 20.04 platform. Furthermore, two NVIDIA GeForce GTX 3060 are used for training and testing. Betas and weight decay are set to (0.9, 0.999) and 0, respectively, in the Adam optimizer. The loss function can be seen in (13) and train the model 600 epochs, the batch size is set to 32. In order to achieve better performance, we initialize the learning rate at 0.0003 and multiply it by 0.9 every 100 epochs. For the traditional methods, we use the MATLAB platform for implementation and comparison.

E. Results on QB Dataset

We first conduct qualitative experiments on the QB dataset to verify the subjective performance of RSANet. The RR test results are shown in Fig. 4. The lower left corner is the zoomed-in area of the red box. It can be clearly seen that the CS-based method shows obvious spatial and spectral distortions, in which the fusion result of CS-based methods distorts the color of the vegetation next to the road into an inhomogeneous sepia color. A similar situation occurs with the three MRA-based methods. In the zoomed-in area, the fusion results of MRA-based methods have blurred road boundaries and are almost integrated with the area outside the roadway, which is not conducive to the subsequent detection and planning of road intersections. As for the DL-based methods, the fusion results of PNN and BDPN are even worse than the traditional methods, with overall severe spatial and spectral distortions and the appearance of colors in the zoomed-in regions that were not present in LRMS and GT. In comparison, the fusion results of PanNet and GPPNN present a slight spectral distortion along with road boundary discontinuities. As regards the spatial resolution, PanNet shows the obvious spatial distortion in the zoomed-in region. The fusion results of GPPNN and SpanConv show the extension of the lawn localization to the pavement region with different degrees of severity, which is a common problem of DL-based methods. The fusion results of TANI and TRRNet have better spectral distributions but introduce some unreasonable impurities at the road edges. In contrast, AWFLN and our method achieve the best fusion results. Moreover, the fusion results of AWFLN have a darker overall color in the lawn region compared to GT. Nevertheless, our method achieves the best spectral preservation, which is almost the same as GT. Furthermore, the last six DL-based methods are even better than PAN in terms of spatial details in subjective visualization, as evidenced by a clearer middle demarcation line of the road in the zoomed-in region, which is more conducive to the subsequent road detection task in this scenario, even though it may not be in line with the real ground situation.

The FR test results are shown in Fig. 5. The lower left corner is the zoomed-in area of the red box. It can be seen that the CS-based methods perform very poorly from a subjective visual perspective, and they show serious distortions on the whole. The fusion results of FS and HPM-R are very close to each other, and they both show a somewhat spectral distortion. In contrast, the performance of PNN, PanNet, and BDPN is

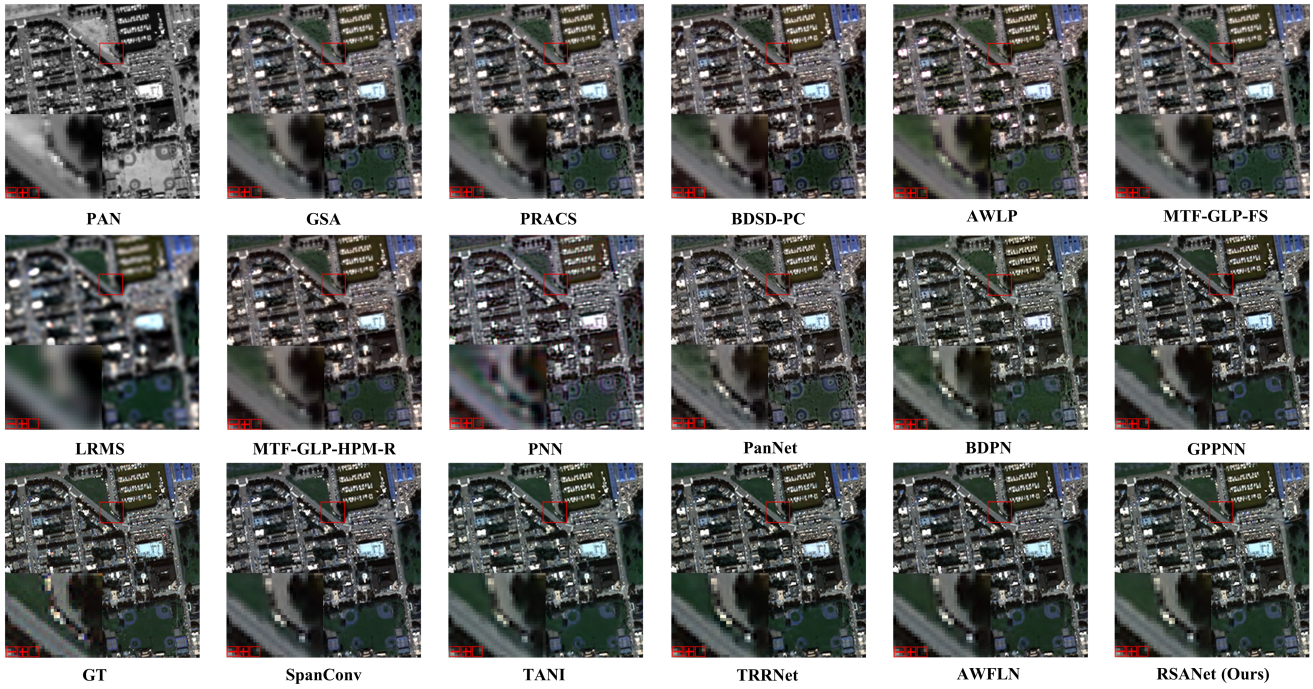

Fig. 4. Qualitative comparison under the RR testing on QB dataset.

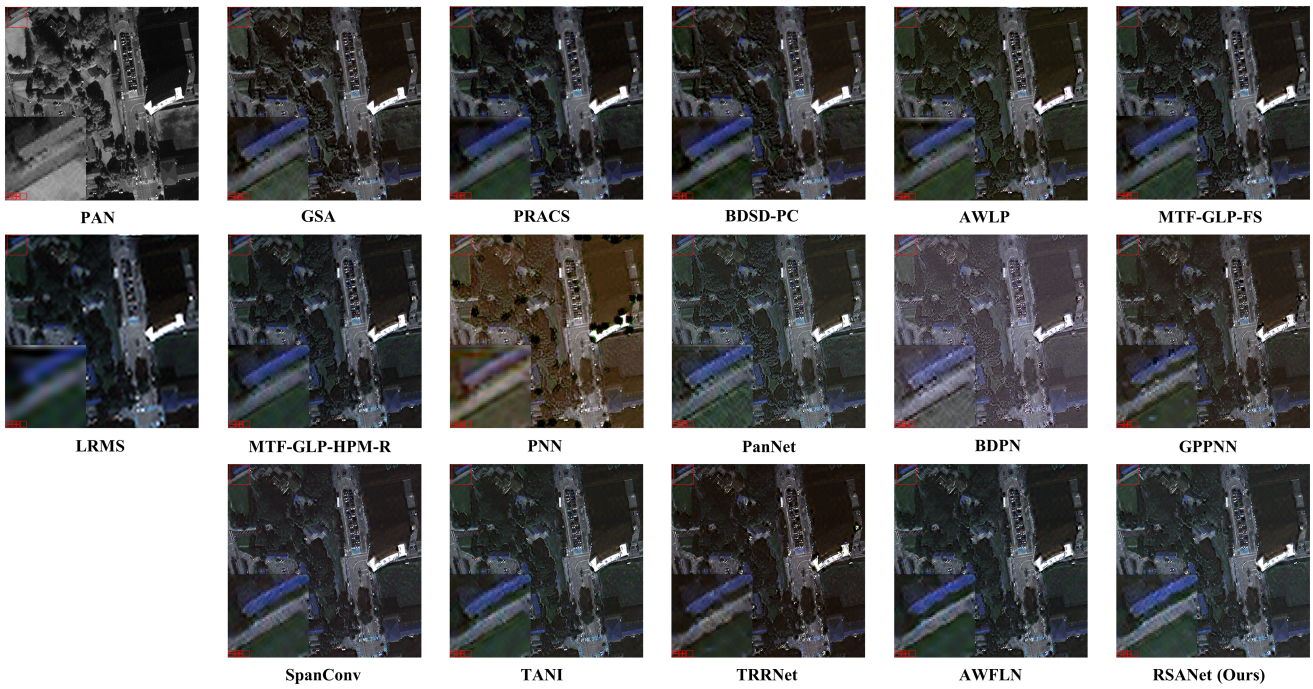

Fig. 5. Qualitative comparison under the FR testing on QB dataset.

the least satisfactory as both exhibit serious spectral distortions. Furthermore, the fusion results of PNN exhibit a general yellowish hue, whereas those of PanNet exhibit a general grayish hue. In conclusion, the latter six DL-based methods all achieve different degrees of balance in spatial and spectral preservation. The fusion results of GPPNN and TRRNet show a slight spatial distortion and two white noises in the zoomed-in region that are not present in the original image pair. The fusion results of TANI show a slight spectral distortion. Furthermore, the fusion results of SpanConv and AWFLN may appear to be inconsistent with the actual situation, as evidenced by the zoomed-in region showing excessive smoothness. Overall, our proposed RSANet achieves the best fusion results, which provides a good basis for subsequent tasks.

We conduct quantitative experiments to further validate the objective performance of RSANet. As given in Table II, the best

TABLE II
QUANTITATIVE COMPARISON OF QUICKBIRD DATASET

Method	Reduced-resolution testing					Full-resolution testing					Params
	SSIM ↑	*SAM* ↓	*CC* ↑	*ERGAS* ↓	*PSNR* ↑	D_λ↓	D_λ^F↓	D_s↓	*QNR*↑	*HQNR*↑*	
GSA	0.857±0.026	8.169±1.717	0.902±0.022	7.699±0.571	32.207±2.295	0.056±0.030	0.234±0.040	0.177±0.027	0.778±0.047	0.631±0.046	/
PRACS	0.836±0.038	8.286±1.779	0.895±0.024	8.457±0.739	31.426±2.477	**0.026±0.012**	0.122±0.023	0.140±0.025	0.838±0.025	0.756±0.036	/
BDSD-PC	0.862±0.029	8.181±1.778	0.903±0.020	7.608±0.574	32.278±2.351	0.026±0.014	0.195±0.033	0.142±0.031	0.836±0.035	0.692±0.046	/
AWLP	0.859±0.032	8.304±1.819	0.905±0.018	7.745±0.701	32.149±2.462	0.047±0.020	0.040±0.010	0.098±0.029	0.861±0.036	0.867±0.033	/
FS	0.864±0.028	7.866±1.628	0.908±0.013	7.445±0.551	32.469±2.264	0.034±0.020	0.045±0.015	0.115±0.023	0.855±0.036	0.845±0.030	/
HPM-R	0.874±0.026	7.835±1.667	0.910±0.025	7.717±1.812	32.600±1.949	0.032±0.021	0.046±0.015	0.105±0.024	0.866±0.032	0.854±0.030	/
PNN	0.861±0.027	8.461±1.535	0.900±0.026	7.591±0.543	32.162±2.115	0.101±0.095	0.183±0.061	0.086±0.044	0.837±0.175	0.749±0.139	0.14M
PanNet	0.880±0.020	7.997±1.450	0.916±0.018	6.999±0.555	32.816±2.080	0.056±0.056	0.076±0.026	0.126±0.036	0.832±0.147	0.807±0.121	0.30M
BDPN	0.927±0.016	6.240±1.165	0.948±0.018	5.473±0.636	34.992±2.150	0.059±0.055	0.072±0.035	0.027±0.015	0.916±0.064	0.903±0.044	5.92M
GPPNN	0.905±0.019	5.860±1.064	0.919±0.014	7.047±0.643	32.947±2.495	0.031±0.027	0.083±0.022	0.029±0.020	0.942±0.042	0.891±0.036	0.24M
SpanConv	0.949±0.008	5.077±0.832	0.966±0.013	4.447±0.259	36.833±1.833	0.027±0.025	0.066±0.015	0.084±0.024	0.892±0.892	0.856±0.033	**0.03M**
TANI	0.935±0.013	5.383±0.901	0.954±0.016	5.222±0.383	35.501±2.042	0.027±0.025	0.073±0.018	0.092±0.026	0.884±0.045	0.842±0.037	0.08M
TRRNet	0.945±0.016	5.015±0.962	0.958±0.021	4.850±0.996	36.199±2.261	0.046±0.039	0.079±0.017	0.030±0.010	0.925±0.041	0.893±0.021	5.21M
AWFLN	0.959±0.006	**4.669±0.786**	0.973±0.008	3.925±0.321	37.842±1.906	0.038±0.024	**0.033±0.015**	0.049±0.023	0.915±0.035	**0.919±0.028**	0.32M
Ours	**0.959±0.005**	4.766±0.797	**0.974±0.009**	**3.869±0.272**	**37.933±1.810**	0.034±0.024	0.062±0.013	**0.025±0.024**	**0.943±0.036**	0.916±0.029	0.18M

* ↑ represents that the larger the value, the better the performance, and ↓ represents that the smaller the value, the better the performance.

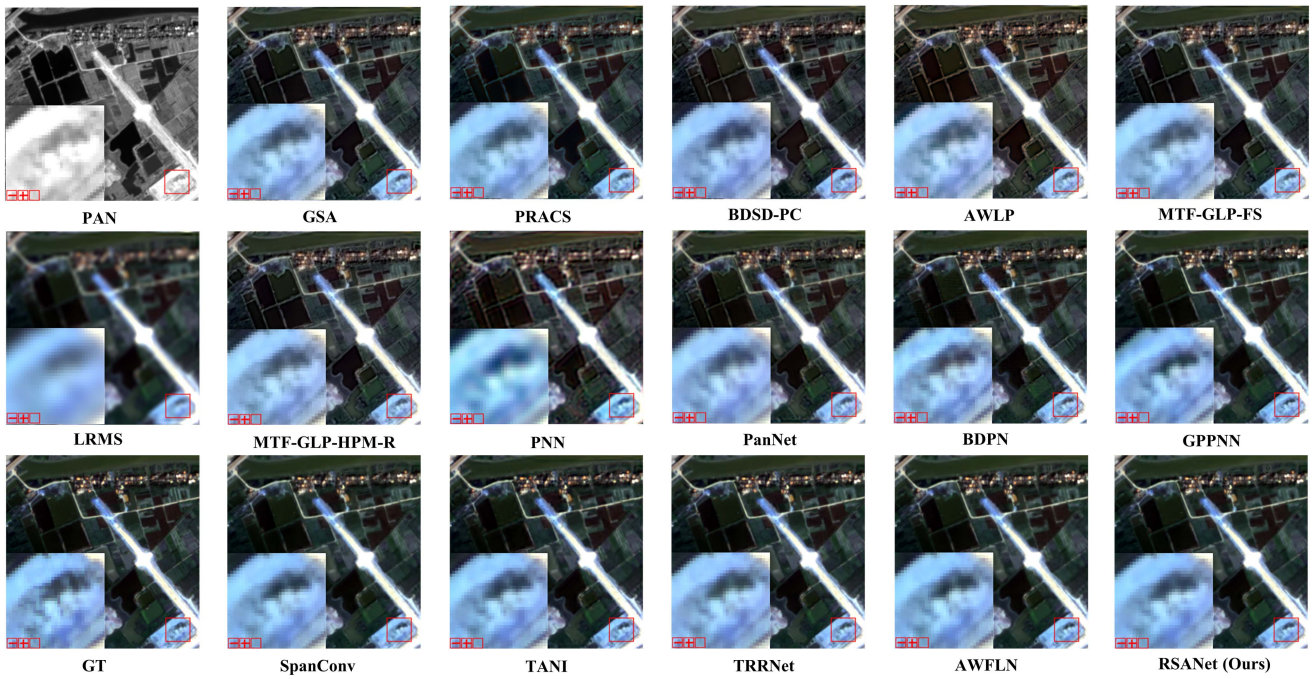

Fig. 6. Qualitative comparison under the RR testing on GF-2 dataset.

values are shown in bold. It can be seen that our proposed method achieves the best scores on six metrics. Specifically, in the RR testing, our method performs best on four metrics, i.e., SSIM, CC, ERGAS, and PSNR, which means that our method best preserves spatial details and spectral information. In the FR testing, we do not perform the best on two spectral distortion indices, i.e., D_λ and D_λ^F, but achieve the best scores on D_s and QNR. QNR has the potential to consider the spatial details injected as spectral distortion during fusion. We can see that the fusion results of AWFLN retain more spectral information when combined with the qualitative analysis, resulting in a slight decrease in our method's performance compared to AWFLN on HQNR. Despite this, the parameters of our method are slightly smaller than those of AWFLN, which demonstrates that our method better balances fusion performance and model parameters.

F. Results on GF-2 Dataset

To further verify the subjective performance of RSANet, we conduct qualitative experiments on the GF-2 dataset. As shown in Fig. 6, in the RR testing, the lower left region shows the zoomed-in state of the purple features in the red box. The results of GSA, PRACS, and BDSD-PC show obvious spectral distortion compared with the GT, which is a common problem of CS-based methods. As shown by the distortion of black features into gray, which can easily mislead the detection of some disasters. For the MRA-based methods, both AWLP and FS show a certain degree of spatial distortion. Their fused results appear vertiginous as a whole, a situation that is not conducive to the detection of urban expansion and thus leads to a greater stress on land resources. With respect to the DL-based methods, the results of PNN show spatial distortions. The results of PanNet and BDPN show slight spatial and spectral distortions, which may be caused by the lack of useful information extracted. The lack of spectral information can cause misunderstandings in feature recognition and is insufficient for natural resource detection. Meanwhile, the fusion results of GPPNN are very close to those of PanNet, and the fusion results of GPPNN are darker. Furthermore, there is some slight spectral distortion in the fusion results of SpanConv and AWFLN, as evidenced

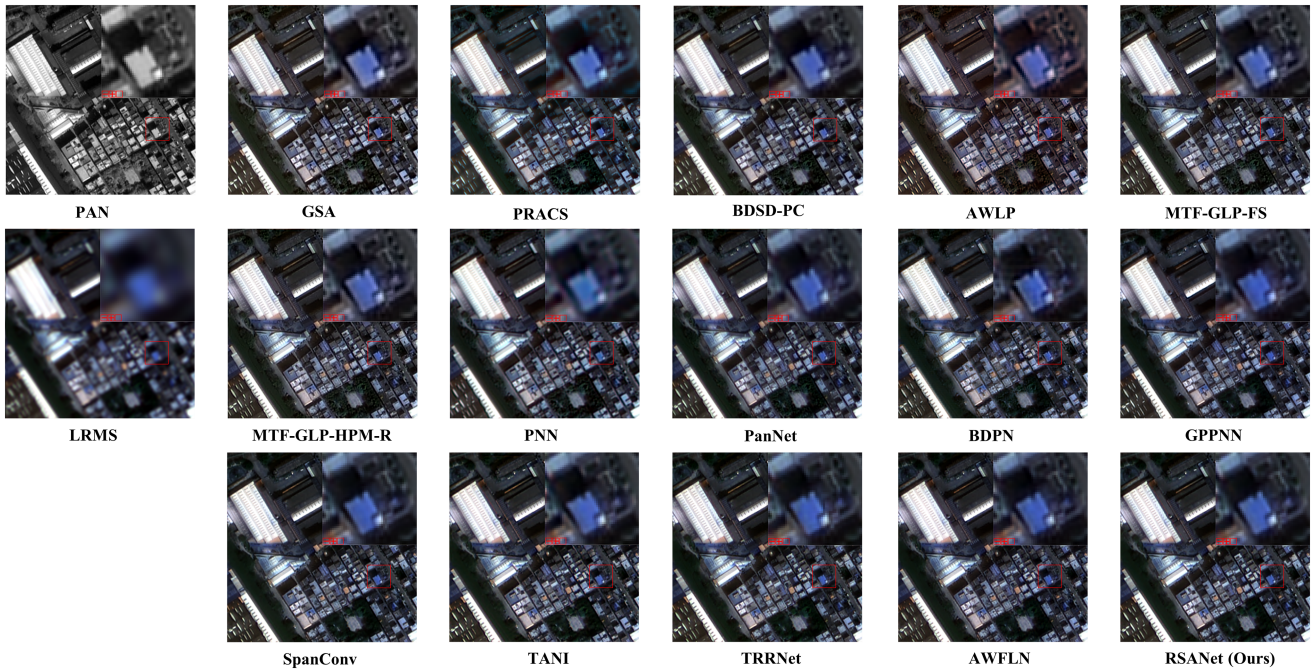

Fig. 7. Qualitative comparison under the FR testing on GF-2 dataset.

TABLE III
QUANTITATIVE COMPARISON OF GAOFEN-2 DATASET

Method	Reduced-resolution testing					Full-resolution testing					Params
	SSIM↑	*SAM*↓	*CC*↑	*ERGAS*↓	*PSNR*↑	D_λ↓	D_S^F↓	D_s↓	*QNR*↑	*HQNR*↑*	
GSA	0.949±0.018	1.734±0.349	0.960±0.020	1.747±0.401	30.819±2.048	0.032±0.019	0.109±0.028	0.070±0.024	0.901±0.039	0.828±0.038	/
PRACS	0.958±0.013	1.714±0.312	0.964±0.020	1.648±0.343	41.302±1.775	0.013±0.009	0.062±0.019	0.047±0.017	0.941±0.023	0.894±0.026	/
BDSD-PC	0.955±0.016	1.724±0.312	0.964±0.018	1.695±0.390	41.046±2.038	0.013±0.010	0.081±0.029	0.056±0.020	0.932±0.026	0.868±0.037	/
AWLP	0.942±0.023	1.968±0.496	0.959±0.018	1.724±0.357	40.560±2.051	0.015±0.014	0.040±0.016	0.049±0.019	0.936±0.030	0.913±0.027	/
FS	0.953±0.016	1.681±0.340	0.965±0.018	1.620±0.353	41.405±1.912	0.024±0.013	0.038±0.013	0.052±0.017	0.925±0.027	0.912±0.021	/
HPM-R	0.954±0.016	1.676±0.347	0.966±0.017	1.620±0.363	41.428±2.003	0.022±0.013	0.036±0.012	0.053±0.018	0.927±0.028	0.913±0.021	/
PNN	0.959±0.011	1.796±0.251	0.964±0.017	1.623±0.257	41.146±1.546	0.021±0.011	0.052±0.022	0.043±0.015	0.937±0.023	0.907±0.028	0.14M
PanNet	0.972±0.007	1.544±0.234	0.976±0.011	1.344±0.194	42.828±1.351	0.012±0.009	0.034±0.026	0.036±0.013	0.953±0.019	0.932±0.030	0.30M
BDPN	0.963±0.014	1.436±0.274	0.970±0.017	1.519±0.368	42.157±2.283	0.012±0.009	0.036±0.020	0.036±0.013	0.952±0.020	0.930±0.024	5.92M
GPPNN	0.971±0.008	1.344±0.233	0.973±0.012	1.414±0.251	42.447±1.735	0.010±0.007	0.043±0.016	0.037±0.013	0.953±0.017	0.921±0.020	0.24M
SpanConv	0.983±0.005	1.012±0.197	0.987±0.007	1.008±0.204	45.551±1.984	0.010±0.007	**0.028±0.010**	0.040±0.014	0.951±0.018	**0.934±0.016**	**0.03M**
TANI	0.978±0.008	1.107±0.212	0.982±0.009	1.149±0.254	44.358±2.079	0.011±0.008	0.041±0.021	0.042±0.016	0.948±0.021	0.919±0.027	0.08M
TRRNet	**0.990±0.003**	0.886±0.124	**0.992±0.004**	0.825±0.115	47.122±1.241	**0.007±0.006**	0.039±0.023	**0.031±0.011**	**0.962±0.015**	0.930±0.024	5.21M
AWFLN	0.989±0.003	0.921±**0.155**	0.991±0.004	0.803±0.119	**47.527±1.458**	0.012±0.009	0.031±0.014	0.040±0.014	0.948±0.019	0.930±0.019	0.32M
Ours	0.989±0.003	**0.885±0.151**	0.991±0.005	**0.801±0.125**	47.475±1.473	0.010±0.008	0.030±0.010	0.038±0.013	0.953±0.016	**0.934±0.016**	0.18M

* ↑ represents that the larger the value, the better the performance, and ↓ represents that the smaller the value, the better the performance.

by the overall mauve coloration of the purple region in the zoomed-in area. In contrast, our method demonstrates the closest deep purple color to GT in the lower left corner. It performs the best overall. However, there is still a certain gap compared to GT, which also alleviates the possible misunderstanding of the subsequent tasks to some extent.

We can see similar results in Fig. 7, where the upper right area is the zoomed-in area of the building in the red box. It can be seen that the fusion results of all the traditional methods exhibit significant spectral distortion, as evidenced by the purple building color being too light in the zoomed-in local area. For DL methods, the fusion results of PNN exhibit severe spatial and spectral distortions. The fusion results of PanNet, TANI, and TRRNet exhibit slight spectral distortions. Furthermore, the fusion results of BDPN, GPPNN, and AWFLN show slight spatial distortion, as evidenced by the introduction of some black noise in the zoomed-in region. In contrast, SpanConv and our method show the best subjective visualization.

Simultaneously, we conduct quantitative experiments on the GF-2 dataset, as given in Table III. In the RR testing, our method achieves the best scores on SAM and ERGAS, which means that our method retains the most spectral information. Meanwhile, our method achieves second place in the other three metrics. In the FR testing, our method achieves second place on QNR and the best score on HQNR as well as SpanConv, which means that the overall quality of the fused images of our method is in the first tier. In terms of parameters, our method has parameters that are 29.5× lower than those of TRRNet and 5.7× higher than those of SpanConv at the same time. To summarize, SpanConv achieves good fusion results despite having the smallest model parameters. However, our method achieves optimal fusion results while limiting the model parameters to a minimal order of magnitude.

G. Results on WV-3 Dataset

We perform the same test on the more challenging 8-band dataset (WV-3 dataset), which contains realistic scenes of buildings, green vegetation, water scenarios, and roads. As shown in Fig. 8, in the RR testing, the lower right area is the zoomed-in area near the road in the red box. It can be seen that

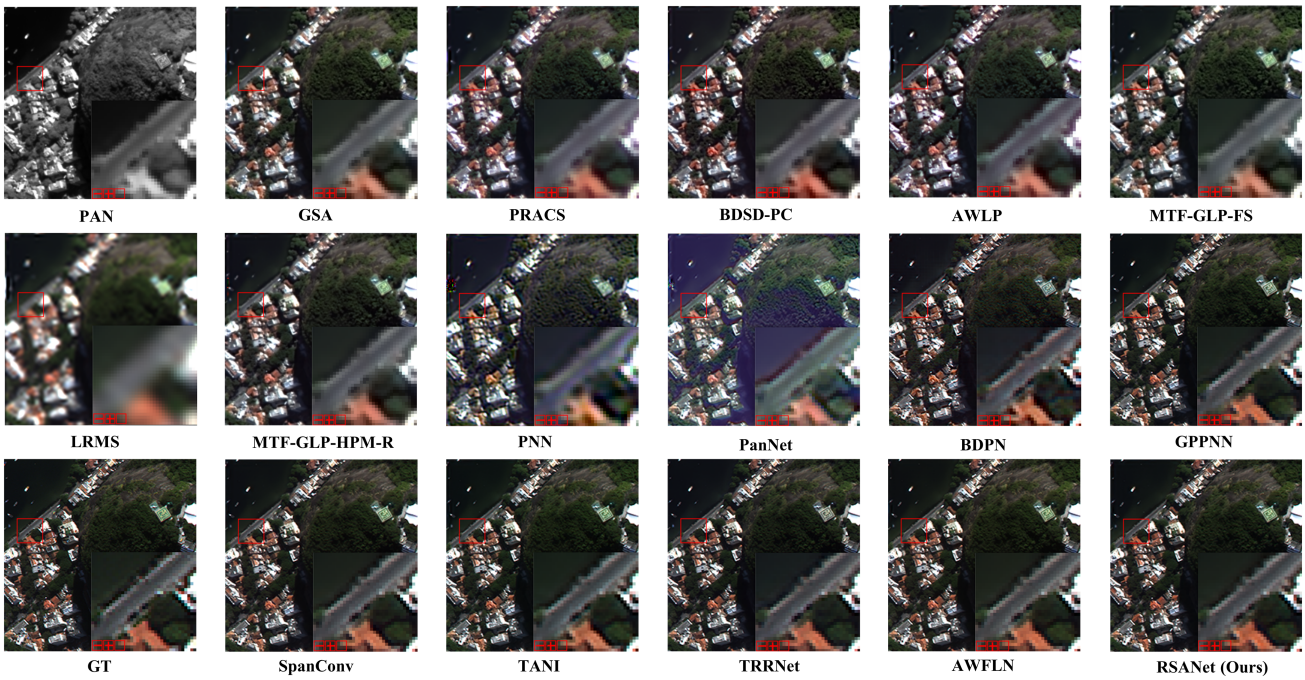

Fig. 8. Qualitative comparison under the RR testing on WV-3 dataset.

the three CS-based methods show serious spatial and spectral distortion, which is manifested in the fact that they distort the roadside water body region in the zoomed-in area into gray. Similar to their performance on other datasets, the MRA-based methods show obvious spatial distortion. Their fusion results are generally blurred, with the fusion results of AWLP accompanied by slight spectral distortion. For the DL-based methods, PNN shows local regional spectral distortion in the zoomed-in region, which is manifested in the middle of the road being distorted into a weird purple color. It is very detrimental to the road screening work. PanNet shows the same problem as PNN but with a different distortion localization. Specifically, the fusion results of PanNet distort the area of the water body next to the road into a purple color, which is very far from the real situation. In contrast, the fusion results of BDPN, GPPNN, SpanConv, and TRRNet do not present much of a problem but show a very slight spectral distortion of the road area on the left in the zoomed-in region. The remaining TANI, AWPLN, and our method show similar blending effects. However, our method has a darker color in the center of the road in the zoomed-in area. Meanwhile, our method shows smoother performance in the road region. In comparison, our method achieves the best subjective effect.

As shown in Fig. 9, in the FR testing, the lower left region is the zoomed-in area of the vehicles in the red box. The CS-based and MRA-based methods demonstrate good subjective visualization but still show slight spatial and spectral distortion in some regions. It is obvious that PRACS, AWLP, and PNN all show spatial distortion, specifically in the form of ghosting of each other's vehicles between neighboring vehicles. These are detrimental to vehicle detection and road planning. As for PanNet, its fusion results show obvious spectral distortion and the overall effect is not as good as the traditional methods. The results of BDPN show spatial distortion. Meanwhile, it can be clearly seen that the fusion results of GPPNN, SpanConv, TANI, and TRRNet show local regional spectral distortion, which is manifested in the fact that they distort the color of some vehicles in the zoomed-in area into other colors. The fusion results of AWPLN exhibit a slight spatial distortion. Compared with PAN and LRMS, our method retains more details of vehicle edges, and the individual vehicle spectral information is not affected by neighboring vehicles.

Furthermore, the results of the quantitative comparison are given in Table IV. In the RR testing, our method has the highest values on three metrics, i.e., SSIM, ERGAS, and PSNR. The remaining two metrics rank behind AWPLN by a narrow margin. In the FR testing, our method achieves the best scores in terms of D_s, QNR, and HQNR. It indicates that despite the fact that the fused image is not the best in terms of spectral information preservation, the more reasonable spatial details contribute to the overall superiority of the fused image. In terms of parameters, it is noteworthy that the model parameters of our method are significantly smaller than those of the second-place finisher, indicating that our method achieves the optimal balance between fusion performance and model parameters even on an eight-band dataset.

V. Ablation Study

In this section, we conduct ablation experiments from three perspectives to verify the performance of RSAM. 1) The two stages of RSAM are treated with ablation experiments. Particularly, in the spatial-spectral similarity extraction stage, the number of recursions (i.e., the value of m) is analyzed. In the self-attention weight generation stage, the effect of the last step of the operation (the last self-attention module) is analyzed. 2) Analyze the effect of the number of RSAMs (i.e., the value

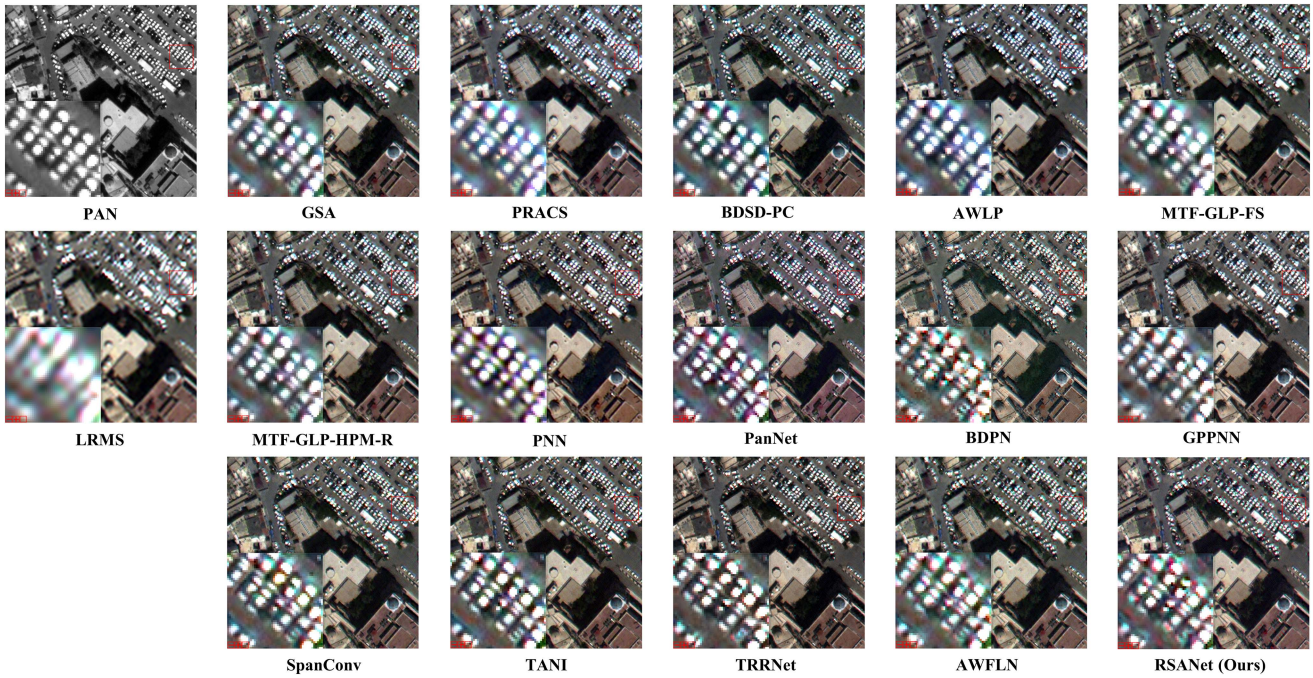

Fig. 9. Qualitative comparison under the FR testing on WV-3 dataset.

TABLE IV
QUANTITATIVE COMPARISON OF WORLDVIEW-3 DATASET

Method	Reduced-resolution testing					Full-resolution testing					Params
	SSIM↑	*SAM*↓	*CC*↑	*ERGAS*↓	*PSNR*↑	D_λ↓	D_λ^F↓	D_s↓	*QNR*↑	*HQNR*↑*	
GSA	0.889±0.031	5.379±1.604	0.933±0.029	4.721±1.422	32.818±2.473	0.023±0.021	0.053±0.023	0.100±0.039	0.880±0.055	0.853±0.055	/
PRACS	0.863±0.034	5.608±1.673	0.920±0.041	5.206±1.469	31.943±2.403	0.014±0.009	0.037±0.013	0.076±0.026	0.912±0.033	0.891±0.035	/
BDSD-PC	0.895±0.029	5.464±1.671	0.935±0.028	4.650±1.427	32.934±2.489	**0.013±0.010**	0.062±0.023	0.091±0.036	0.897±0.043	0.853±0.051	/
AWLP	0.895±0.022	5.276±1.365	0.933±0.027	4.697±1.329	32.672±2.648	0.022±0.016	**0.016±0.008**	0.076±0.031	0.904±0.043	0.909±0.037	
FS	0.891±0.030	5.323±1.611	0.936±0.029	4.645±1.406	32.954±2.454	0.020±0.017	0.020±0.008	0.085±0.031	0.897±0.043	0.897±0.035	
HPM-R	0.893±0.030	5.369±1.601	0.931±0.026	5.611±3.102	32.751±2.536	0.021±0.017	0.021±0.008	0.086±0.031	0.896±0.044	0.896±0.034	/
PNN	0.861±0.028	6.543±1.549	0.904±0.060	5.240±1.326	31.328±2.022	0.030±0.014	0.051±0.020	0.083±0.030	0.889±0.041	0.871±0.043	0.18M
PanNet	0.894±0.018	6.940±1.386	0.928±0.063	5.234±1.513	32.079±1.551	0.025±0.019	0.063±0.024	0.093±0.035	0.885±0.049	0.850±0.048	0.31M
BDPN	0.943±0.013	4.448±0.963	0.958±0.042	3.371±0.786	35.363±2.125	0.024±0.016	0.024±0.009	0.053±0.017	0.923±0.028	0.924±0.021	5.94M
GPPNN	0.969±0.009	3.309±0.626	0.978±0.016	2.483±0.525	37.896±2.652	0.013±0.010	0.042±0.016	0.062±0.024	0.926±0.034	0.899±0.033	0.48M
SpanConv	0.966±0.010	3.505±0.670	0.976±0.017	2.590±0.539	37.578±2.749	0.015±0.012	0.026±0.009	0.076±0.025	0.911±0.038	0.900±0.031	**0.03M**
TANI	0.964±0.010	3.605±0.670	0.974±0.020	2.687±0.563	37.300±2.608	0.016±0.015	0.035±0.014	0.085±0.032	0.901±0.043	0.883±0.042	0.09M
TRRNet	0.971±0.008	3.263±0.561	0.980±0.016	2.364±0.513	38.244±2.469	0.017±0.015	0.036±0.014	0.074±0.024	0.911±0.035	0.893±0.032	5.23M
AWFLN	0.972±0.009	**3.071±0.568**	**0.981±0.014**	2.297±0.528	38.479±2.626	0.016±0.014	0.020±0.010	0.063±0.023	0.922±0.033	0.918±0.029	0.34M
Ours	**0.972±0.008**	3.125±0.546	0.980±0.017	**2.290±0.495**	**38.485±2.558**	0.020±0.015	0.023±0.010	**0.050±0.023**	**0.931±0.027**	**0.928±0.024**	0.20M

* ↑ represents that the larger the value, the better the performance, and ↓ represents that the smaller the value, the better the performance.

of n) on the model performance and parameters. 3) Compare the proposed RSAM with other attentional models in the same model framework. It is worth noting that the number of recursions in the first stage of the proposed RSAM is 3, while the number of RSAMs in the proposed RSANet is 8 (i.e., $m = 3$, $n = 8$). In order to better validate the performance of RSAM, we perform all ablation experiments on the more challenging eight-band dataset (WV-3 dataset).

A. Number of Recursions and the Effect of the Last Step

In the spatial-spectral similarity extraction stage, we conduct experiments to verify the effect of the number of recursions on the similarity extraction capability. The results of the qualitative experiments are shown in Fig. 10, with the red box enlarged in the lower left corner. The RR testing shows that the fusion results exhibit a slight spectral distortion when $m = 0$, as shown by the dark green color of the vegetation in the lower right corner of the image and the gray-green color in the GT. As the number of recursions increases, the spectral distribution of the fused image becomes increasingly similar to that of GT. At an m value of 4, the fusion results exhibit excessive spatial detail, resulting in an unreasonable spectral distribution, as evidenced by the yellow hue of the vegetation in the upper right corner of the pool. In the FR testing, it can be seen that the spatial details and spectral distribution of the fused image get closer to the original image pair as the number of recursions increases. The quantitative experiments are presented in Table V, wherein the proposed RSANet achieves the best values on all five referenced metrics in the RR testing. In the FR testing, RSANet excels on D_λ and QNR, while RSANet ($m = 4$) excels on D_λ^F and HQNR. This is attributed to the fact that the details injected during fusion may be regarded as spectral distortions in QNR, which is in accordance with the findings of the qualitative analysis. Combining the results of qualitative and quantitative experiments, the proposed RSAM can achieve the best performance.

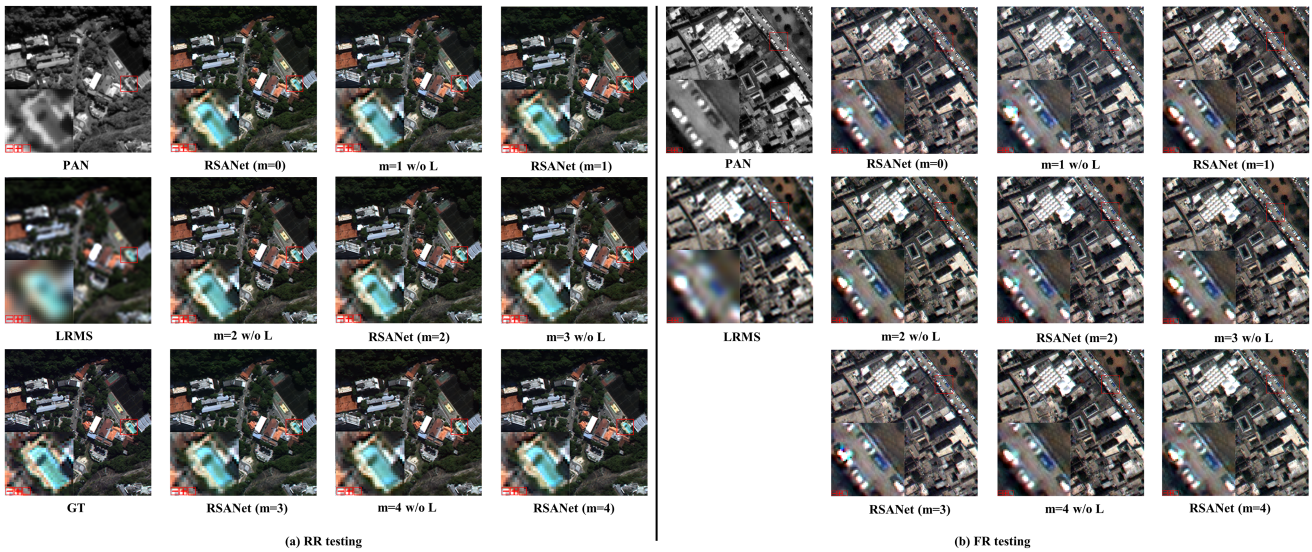

Fig. 10. Qualitative comparison of the number of recursions and the effect of last step. (a) RR testing. (b) FR testing.

TABLE V
QUANTITATIVE COMPARISON OF THE NUMBER OF RECURSIONS AND THE EFFECT OF THE LAST STEP

Method	Reduced-resolution testing					Full-resolution testing					Params
	SSIM↑	*SAM*↓	*CC*↑	*ERGAS*↓	*PSNR*↑	D_λ↓	D_s^F↓	D_s↓	*QNR*↑	*HQNR*↑*	
RSANet(m=0)	0.970±0.009	3.240±0.581	0.979±0.018	2.401±0.552	38.166±2.450	0.031±0.050	0.026±0.015	0.060±0.042	0.913±0.079	0.917±0.051	0.15M
m=1 w/o L	0.965±0.009	3.581±0.667	0.975±0.020	2.620±0.551	37.394±2.587	0.095±0.183	0.030±0.015	0.147±0.201	0.808±0.256	0.830±0.200	0.15M
RSANet(m=1)	0.971±0.008	3.189±0.571	0.980±0.017	2.331±0.485	38.384±2.500	0.042±0.038	0.028±0.013	**0.048±0.033**	0.912±0.053	0.926±0.036	0.16M
m=2 w/o L	0.966±0.009	3.436±0.619	0.977±0.019	2.536±0.570	37.684±2.413	0.089±0.158	0.034±0.022	0.143±0.168	0.806±0.235	0.831±0.171	0.16M
RSANet(m=2)	0.971±0.007	3.200±0.558	0.980±0.016	2.319±0.497	38.424±2.480	0.021±0.016	0.023±0.009	0.056±0.021	0.924±0.026	0.923±0.023	0.18M
m=3 w/o L	0.963±0.012	3.474±0.624	0.974±0.022	2.680±0.649	37.229±2.296	0.044±0.094	0.028±0.012	0.113±0.131	0.861±0.170	0.863±0.132	0.18M
RSANet	**0.972±0.008**	**3.125±0.546**	**0.980±0.017**	**2.290±0.495**	**38.485±2.558**	**0.020±0.015**	0.023±0.010	0.050±0.023	**0.931±0.027**	0.928±0.024	0.20M
m=4 w/o L	0.968±0.009	3.393±0.608	0.977±0.019	2.511±0.594	37.790±2.441	0.039±0.079	0.025±0.012	0.102±0.122	0.872±0.155	0.876±0.124	0.20M
RSANet(m=4)	0.969±0.009	3.293±0.638	0.978±0.021	2.412±0.527	38.123±2.271	0.021±0.012	**0.020±0.012**	0.053±0.019	0.927±0.025	**0.928±0.022**	0.22M

* ↑ represents that the larger the value, the better the performance, and ↓ represents that the smaller the value, the better the performance.

When the number of recursions is three, RSAM is capable of achieving a satisfactory balance of the similarity between the extracted individual spectral bands and the similarity between spatial pixel points. Meanwhile, the results show more spatial details when the number of recursions is four, which breaks the balance between spectral and spatial information and leads to unreasonable spectral distributions.

On this basis, we fix the number of recursions and then remove the self-attention module in the self-attention weight generation stage to verify the criticality of the last step operation. The results of the qualitative experiments are shown in Fig. 10, where w/o L denotes the removal of the last step operation. In the RR testing, the spatial details and spectral distributions in the fused image do not change as the number of recursions in the first stage when the last step is removed. This is evident from the fact that the swimming pool portion in the lower left corner exhibits obvious spectral distortion. In FR testing, the white vehicle in the enlarged region after removing the last step shows obvious spatial distortion when $m = 1$. Meanwhile, when $m = 2$, $m = 3$, and $m = 4$, the fusion results after removing the last step show obvious spectral distortion, which is specifically manifested in the appearance of a yellow halo around the white vehicle in the enlarged region. This is inconsistent with what is presented by PAN and LRMS. As given in Table V, the quantitative results of removing the last step are much smaller than those of RSANet under the same parameters. The reason for this is that the proposed RSAM adopts a G2L strategy, and in the spatial-spectral similarity extraction stage, we adequately extract the similarity between each spectral band and each spatial pixel point to establish global dependencies by the RSAMs. In the self-attention weight generation stage, we feed the global similarity information extracted in the first stage into the last self-attention module (also referred to as the last step) in blocks to establish dependencies between each localization. Subsequently, we accurately represent the importance of the important features in each localization through the sigmoid activation function and finally obtain the self-attention weights of each localization. Finally, we get the self-attention weight of each localization. Therefore, the last step is imperative.

B. Number of RSAMs

As the proposed RSANet is primarily based on RSAMs, the number of RSAMs is an important factor that has an impact on both the fusion performance and model parameters. We test five designs of RSANet to evaluate the impact of the number of RSAMs. The results of the qualitative experiments are shown in Fig. 11, where n is the number of RSAMs. In the RR testing,

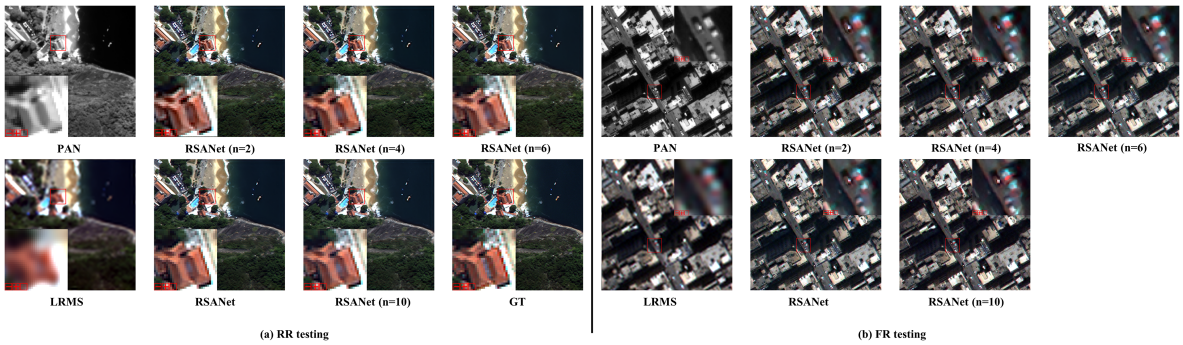

Fig. 11. Qualitative comparison of the number of RSAMs. (a) RR testing. (b) FR testing.

TABLE VI
QUANTITATIVE COMPARISON OF THE NUMBER OF RSAMS

Method	Reduced-resolution testing					Full-resolution testing					Params
	SSIM↑	*SAM*↓	*CC*↑	*ERGAS*↓	*PSNR*↑	D_λ↓	D_λ^F↓	D_s↓	*QNR*↑	*HQNR*↑*	
RSANet(n=2)	0.957±0.011	3.972±0.734	0.970±0.023	2.909±0.675	36.569±2.516	0.015±0.013	0.037±0.012	0.073±0.027	0.913±0.035	0.893±0.034	**0.03M**
RSANet(n=4)	0.967±0.009	3.381±0.607	0.977±0.020	2.498±0.590	37.807±2.411	**0.015±0.013**	0.026±0.011	0.071±0.026	0.916±0.035	0.906±0.033	0.09M
RSANet(n=6)	0.969±0.009	3.326±0.620	0.978±0.019	2.454±0.573	37.992±2.385	0.015±0.012	**0.022±0.009**	0.064±0.021	0.922±0.030	0.916±0.026	0.14M
RSANet	**0.972±0.008**	**3.125±0.546**	**0.980±0.017**	**2.290±0.495**	**38.485±2.558**	0.020±0.015	0.023±0.010	**0.050±0.023**	**0.931±0.027**	0.928±0.024	0.20M
RSANet(n=10)	0.971±0.008	3.204±0.603	0.979±0.019	2.335±0.488	38.367±2.357	0.021±0.018	0.022±0.009	0.050±0.023	0.930±0.028	**0.929±0.023**	0.26M

* ↑ represents that the larger the value, the better the performance, and ↓ represents that the smaller the value, the better the performance.

the bottom left corner shows the zoomed-in effect of the red box. When the value of n is between 2 and 6, the fusion results show obvious spectral distortion, which is evident in the overall light color of the orange building. In contrast, the proposed RSANet has a spectral distribution similar to that of GT. In the FR testing, the top right corner is the zoomed-in effect of the red box. It is evident that as the number of RSAMs increases, the spatial details and spectral information on the wagon become increasingly detailed. However, when $n = 10$, the increased spatial details result in unreasonable spectral distributions, as evident by the distortion of some of the color information on the sedan to the road surface.

We also conduct quantitative experiments, and the results are given in Table VI. In the RR testing, as the number of RSAMs increases, the fusion results of RSANet become better in terms of the average values on the five referenced metrics. The scores on the five metrics decrease slightly with $n = 10$ but more stable quantitative results are obtained, as evidenced by the smaller standard deviations on the three metrics of SSIM, ERGAS, and PSNR. In the FR testing, as the number of RSAMs increases, the performance of the fused images on the nonreference metrics improves. When $n = 10$, the value of QNR decreases while the value of HQNR increases. This is due to the unreasonable spectral distribution caused by injecting some spatial details into the fused image. In general, the number of RSAMs exhibits a positive correlation with the performance of fusion. Furthermore, the model parameters increase on average by 0.03M for each additional RSAM. Therefore, the proposed RSANet (i.e., $n = 8$) achieves the best balance between fusion performance and model parameters.

C. Compare With Other Attention Mechanisms

To further validate the performance of RSAM, we compare it objectively against other attention modules within the same modeling framework. The model framework follows the structure shown in Fig. 3 ($m = 8$). Furthermore, we select four attention modules, mainly comprising two generalized attention modules (CBAM and GAM) and two image fusion-based attention modules (IIM and SWM), of which IIM and SWM are the basic modules used in the comparison methods TANI and TRRNet, respectively.

The results of the qualitative experiments are shown in Fig. 12, with the effect enlarged by the red box located in the lower left corner. In the RR testing, the CBAM- and IIM-based fusion results show obvious spatial and spectral distortion. Meanwhile, a slight spectral distortion can be seen in the zoomed-in area of the GAM-based fusion results. In the lower right corner of the zoomed-in area of the SWM-based fusion results, purple pixel points are observed between the vegetation gaps, whereas they are not observed in GT. Although our proposed method still has some gaps compared to GT, the overall spatial details and spectral distributions of the fused images are roughly the same as GT. In the FR testing, the fusion results based on CBAM and IIM show significant spectral distortion. The zoomed-in region shows that our proposed method remains largely consistent with the original image pair in terms of spatial details and spectral information. The results of the qualitative experiments are given in Table VII. In the RR testing, our method achieves the best mean values on five metrics and the smallest standard deviations on four metrics. In the FR testing, despite our method failing to achieve optimal results on the two spectral distortion metrics, it achieves superior results on the spatial distortion metrics, which are reasonably spatially detailed to drive the fused images to achieve better spectral distributions. As a result, our method achieves optimal results on the two metrics QNR and HQNR. Combining both the qualitative and quantitative experimental results, it can be concluded that the proposed RSAM exhibits better performance compared to other attention modules within a specific model framework.

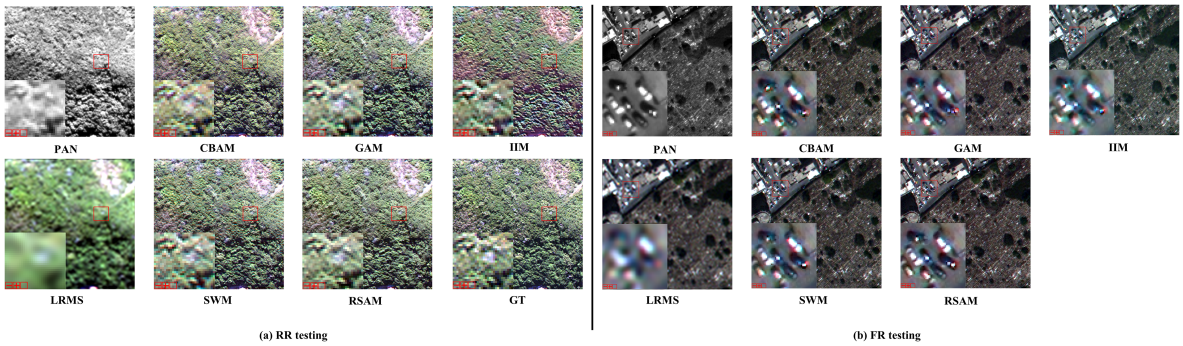

Fig. 12. Qualitative comparison with other attention mechanism. (a) RR testing. (b) FR testing.

TABLE VII
QUANTITATIVE COMPARISON WITH OTHER ATTENTION MECHANISMS

Method	Reduced-resolution testing					Full-resolution testing					Params
	SSIM↑	*SAM*↓	*CC*↑	*ERGAS*↓	*PSNR*↑	D_λ↓	D_s^F↓	D_s↓	*QNR*↑	*HQNR*↑*	
CBAM	0.959±0.010	3.642±0.771	0.971±0.026	2.851±0.813	36.792±1.907	0.022±0.015	0.024±0.009	0.061±0.021	0.919±0.029	0.917±0.025	0.24M
GAM	0.969±0.008	3.335±0.600	0.978±0.019	2.448±0.578	37.990±2.405	0.022±0.021	**0.021±0.009**	0.059±0.047	0.921±0058	0.921±0.048	0.32M
IIM(TANI)	0.951±0.011	4.308±0.811	0.964±0.033	3.141±0.776	35.879±2.154	0.020±0.018	0.043±0.015	0.084±0.028	0.898±0.040	0.877±0.037	0.10M
SWM(TRRNet)	0.968±0.009	3.337±0.625	0.978±0.019	2.489±0.564	37.903±2.407	**0.016±0.013**	0.029±0.009	0.072±0.021	0.914±0.031	0.902±0.026	0.61M
RSAM	**0.972±0.008**	**3.125±0.546**	**0.980±0.017**	**2.290±0.495**	**38.485±2.558**	0.020±0.015	0.023±0.010	**0.050±0.023**	**0.931±0.027**	**0.928±0.024**	0.20M

* ↑ represents that the larger the value, the better the performance, and ↓ represents that the smaller the value, the better the performance.

VI. DISCUSSION

Considering the performance of the fusion effect and the limitations of hardware specifications in practical application scenarios, we propose a novel RSAM to effectively balance fusion performance and model parameters. Our method is able to preserve more spatial details and spectral information than traditional image fusion methods, which is mainly due to the fact that traditional methods can only focus on either the spatial domain or the spectral domain. In contrast, the proposed RSAM employs a G2L way to capture the global interdependencies of two distinct local locations in the feature map, which can consider both spatial and spectral information simultaneously and can establish global and local dependencies to better extract and retain important information. Our method is more efficient than traditional CNN fusion networks because conventional convolution requires many layers to be stacked to extend the local receptive field to the global area, whereas our method achieves the best experimental results in comparative experiments by stacking only eight RSAMs. Compared with other attention modules, our method is capable of establishing the global and local dependence between spectral bands and spatial pixel points while focusing on more mutual information between spectral and spatial dimensions, which allows for more spatial details and spectral information to be preserved under complex remote sensing images. Overall, our method has a relatively small number of parameters, which contributes to superior fusion results. However, this compromises the timeliness of the model in part due to the large number of matrix multiplication operations required for attention in the module. In the subsequent work, we intend to enhance the time efficiency of the model to better cater to the requirements of Earth observation.

VII. CONCLUSION

In this work, we design a novel RSAM, which employs the G2L strategy to establish global dependencies by calculating the similarity between individual spectral bands and between spatial pixel points. Then, establishes local representations with the global information, which is capable of capturing features of different resolutions and better recovering spatial details in multispectral images. Furthermore, we design the corresponding residual block RSARB by RSAM and constitute RSANet. RSANet can achieve the best fusion effect with a small number of parameters on three publicly available satellite datasets. It shows that our method has a strong feature learning ability for important information, which provides basic technical support for observing global resource and environmental conditions, thus promoting sustainable human development.

Despite the fact that the proposed method achieves a better balance between fusion performance and model parameters, there remain some noteworthy concerns that require further investigation. In our experiments, we use Wald's protocol for model training, which is not able to fully simulate the physical degradation process in the real world. It is necessary to further study the degradation process in the real world and then simulate it in order to improve the performance of supervised learning. In the meantime, we intend to utilize the self-supervised learning method to mine the information from remote sensing images for training on real data. Furthermore, since fusion methods based on DL are constrained by various aspects in application scenarios such as on-board processing, we also intend to develop more lightweight models to provide the possibility of on-board fusion for DL as well as the prerequisites for subsequent detection tasks.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and members of the editorial team for their comments and suggestions.

REFERENCES

[1] A.-V. Emilien, C. Thomas, and H. Thomas, "UAV & satellite synergies for optical remote sensing applications: A literature review," *Sci. Remote Sens.*, vol. 3, Jun. 2021, Art. no. 100019, doi: 10.1016/j.srs.2021.100019.

[2] Y. Xu et al., "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019, doi: 10.1109/jstars.2019.2911113.

[3] Z. Tian, R. Zhan, W. Wang, Z. He, J. Zhang, and Z. Zhuang, "Object detection in optical remote sensing images by integrating object-to-object relationships," *Remote Sens. Lett.*, vol. 11, no. 5, pp. 416–425, Feb. 2020, doi: 10.1080/2150704x.2020.1722330.

[4] N. Long Nguyen, J. Anger, A. Davy, P. Arias, and G. Facciolo, "L1BSR: Exploiting detector overlap for self-supervised single-image super-resolution of sentinel-2 L1B imagery," Cornell Univ., Ithaca, NY, USA, Apr. 2023, doi: 10.48550/arxiv.2304.06871.

[5] X. Tian, Y. Chen, C. Yang, X. Gao, and J. Ma, "A variational pansharpening method based on gradient sparse representation," *IEEE Signal Process. Lett.*, vol. 27, pp. 1180–1184, 2020, doi: 10.1109/LSP.2020.3007325.

[6] G. Vivone et al., "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021, doi: 10.1109/MGRS.2020.3019315.

[7] J.-L. Xiao, T.-Z. Huang, L.-J. Deng, Z.-C. Wu, and G. Vivone, "A new context-aware details injection fidelity with adaptive coefficients estimation for variational pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5408015, doi: 10.1109/TGRS.2022.3154480.

[8] H. Lu, Y. Yang, S. Huang, W. Tu, and W. Wan, "A unified pansharpening model based on band-adaptive gradient and detail correction," *IEEE Trans. Image Process.*, vol. 31, pp. 918–933, 2022, doi: 10.1109/TIP.2021.3137020.

[9] Y. Gao, M. Zhang, W. Li, X. Song, X. Jiang, and Y. Ma, "Adversarial complementary learning for multisource remote sensing classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5505613, doi: 10.1109/tgrs.2023.3255880.

[10] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and LiDAR data classification based on structural optimization transmission," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3153–3164, May 2023, doi: 10.1109/tcyb.2022.3169773.

[11] M. Zhou, J. Huang, D. Hong, F. Zhang, C. Li, and J. Chanussot, "Rethinking pan-sharpening in closed-loop regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/tnnls.2023.3279931.

[12] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021, doi: 10.1109/TGRS.2020.3016820.

[13] D. Hong, J. Yao, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," Cornell Univ., Ithaca, NY, USA, May 2022, doi: 10.48550/arxiv.2205.03742.

[14] Z. Zhang, W. Lü, X. Shao, G. Xie, C. Liu, and M. Xu, "Task-driven on-board real-time panchromatic multispectral fusion processing approach for high-resolution optical remote sensing satellite," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 7636–7661, 2023, doi: 10.1109/jstars.2023.3305231.

[15] M. Wang, Z. Zhang, Y. Zhu, Z. Dong, and Y. Li, "Embedded GPU implementation of sensor correction for on-board real-time stream computing of high-resolution optical satellite imagery," *J. Real-Time Image Process.*, vol. 15, no. 3, pp. 565–581, 2018, doi: 10.1007/s11554-017-0741-0.

[16] Z. Z. M. Wang, "Stream-computing of high accuracy on-board real-time cloud detection for high resolution optical satellite imagery," *J. Geodesy Geoinf. Sci.*, vol. 2, no. 2, pp. 50–59, Mar. 2020, doi: 10.11947/j.JGGS.2019.0206.

[17] Z. Zhang, Z. Qu, S. Liu, D. Li, J. Cao, and G. Xie, "Expandable on-board real-time edge computing architecture for Luojia3 intelligent remote sensing satellite," *Remote Sens.*, vol. 14, no. 15, Jan. 2022, Art. no. 3596, doi: 10.3390/rs14153596.

[18] S. Xiang, Q. Liang, and P. Tang, "Task-oriented compression framework for remote sensing satellite data transmission," *IEEE Trans. Ind. Inform.*, to be published, doi: 10.1109/TII.2023.3309030.

[19] S. Xiang, Q. Liang, and L. Fang, "Discrete wavelet transform-based Gaussian mixture model for remote sensing image compression," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3000112.

[20] Y. Gao, M. Zhang, J. Wang, and W. Li, "Cross-scale mixing attention for Multisource remote sensing data fusion and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507815, doi: 10.1109/tgrs.2023.3263362.

[21] S. Xiang and Q. Liang, "Remote sensing image compression with long-range convolution and improved non-local attention model," *Signal Process*, vol. 209, 2023, Art. no. 109005.

[22] S. Xiang, M. Wang, X. Jiang, G. Xie, Z. Zhang, and P. Tang, "Dual-task semantic change detection for remote sensing images using the generative change field module," *Remote Sens.*, vol. 13, no. 16, Aug. 2021, Art. no. 3336, doi: 10.3390/rs13163336.

[23] Z. Zhang, W. Xia, G. Xie, and S. Xiang, "Fast opium poppy detection in unmanned aerial vehicle (UAV) imagery based on deep neural network," *Drones*, vol. 7, no. 9, Sep. 2023, Art. no. 559, doi: 10.3390/drones7090559.

[24] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415, doi: 10.1109/tgrs.2023.3284671.

[25] W. J. Carper, "The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data," *Photogramm. Eng. Remote Sens.*, vol. 56, no. 4, pp. 457–467, Jan. 1990.

[26] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. Channel ratio and 'chromaticity' transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, Aug. 1987, doi: 10.1016/0034-4257(87)90088-5.

[27] S. Lolli, L. Alparone, A. Garzelli, and G. Vivone, "Haze correction for contrast-based multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2255–2259, Dec. 2017, doi: 10.1109/lgrs.2017.2761021.

[28] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS+pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007, doi: 10.1109/tgrs.2007.901007.

[29] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228–236, Jan. 2008, doi: 10.1109/tgrs.2007.907604.

[30] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2011, doi: 10.1109/tgrs.2010.2051674.

[31] X. Otazu, M. Gonzalez-Audicana, O. Fors, and J. Nunez, "Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2376–2385, Oct. 2005, doi: 10.1109/tgrs.2005.856106.

[32] L. Alparone, A. Garzelli, and G. Vivone, "Intersensor statistical matching for pansharpening: Theoretical issues and practical solutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4682–4695, Aug. 2017, doi: 10.1109/tgrs.2017.2697943.

[33] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Dec. 2002, doi: 10.1109/tgrs.2002.803623.

[34] H. A. Aly and G. Sharma, "A regularized model-based optimization framework for pan-sharpening," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2596–2608, Jun. 2014, doi: 10.1109/tip.2014.2316641.

[35] L.-J. Deng, M. Feng, and X.-C. Tai, "The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior," *Inf. Fusion*, vol. 52, pp. 76–89, Dec. 2019, doi: 10.1016/j.inffus.2018.11.014.

[36] C. Ballester, V. Caselles, L. Igual, J. O. Verdera, and B. Rougé, "A variational model for P+XS image fusion," *Int. J. Comput. Vis.*, vol. 69, no. 1, pp. 43–58, Apr. 2006, doi: 10.1007/s11263-006-6852-x.

[37] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011, doi: 10.1109/tgrs.2010.2067219.

[38] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, Jul. 2016, Art. no. 594, doi: 10.3390/rs8070594.

[39] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5449–5457, doi: 10.1109/iccv.2017.193.

[40] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, and C. Zhang, "Deep gradient projection networks for pan-sharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1366–1375, doi: 10.1109/cvpr46437.2021.00142.
[41] Z.-X. Chen et al., "SpanConv: A new convolution via spanning kernel space for lightweight pansharpening," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022, pp. 1–7, doi: 10.24963/ijcai.2022/118.
[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/cvpr.2018.00745.
[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
[44] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lü, "TAM: Temporal adaptive module for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 13708–13718, doi: 10.1109/iccv48922.2021.01345.
[45] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," Dec. 10, 2021, *arXiv:2112.05561*.
[46] W. Diao, F. Zhang, H. Wang, J. Sun, and K. Zhang, "Pansharpening via triplet attention network with information interaction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3576–3588, 2022, doi: 10.1109/jstars.2022.3171423.
[47] K. Zhang, Z. Li, F. Zhang, W. Wan, and J. Sun, "Pan-Sharpening Based on Transformer With Redundancy Reduction," *IEEE Geosci. Remote Sens.*, vol. 19, pp. 1–5, Jan. 2022, doi: 10.1109/lgrs.2022.3186985.
[48] H. Lu et al., "AWFLN: An adaptive weighted feature learning network for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5400815, doi: https://doi.org/10.1109/tgrs.2023.3241643.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778, doi: 10.1109/cvpr.2016.90.
[50] H. Chen, J. Gu, and Z. Zhang, "Attention in attention network for image super-resolution," 2021, *arXiv:2104.09497*.
[51] L. J. Deng et al., "CNN-based remote sensing pan-sharpening: A critical review," *J. Image Graph.*, vol. 28, no. 1, pp. 57–79, 2023.
[52] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997, doi: 10.1016/S0924-2716(97)00008-7.
[53] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/tip.2003.819861.
[54] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, Oct. 2007, doi: 10.1109/tgrs.2007.904923.
[55] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, Jan. 1998, doi: 10.1080/014311698215973.
[56] L. Wald, "Data fusion: Definitions and architectures: Fusion of images of different spatial resolutions," Presses des MINES, 2002.
[57] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?," in *Proc. 4th Int. Workshop Qual. Multimedia Experience*, Jul. 2012, pp. 37–38, doi: 10.1109/qomex.2012.6263880.
[58] L. Alparone et al., "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, 2008, doi: 10.14358/pers.74.2.193.
[59] A. Arienzo, G. Vivone, A. Garzelli, L. Alparone, and J. Chanussot, "Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 168–201, Sep. 2022, doi: 10.1109/mgrs.2022.3170092.
[60] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421–6433, Sep. 2019, doi: 10.1109/tgrs.2019.2906073.
[61] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018, doi: 10.1109/tip.2018.2819501.
[62] Y. Zhang, L. Chi, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549–5563, Aug. 2019, doi: 10.1109/tgrs.2019.2900419.

Chuang Liu received the B.Sc. degree in computer science and technology from the Wuhan University of Engineering Science, Wuhan, China, in 2022. He is currently working toward the M.Eng. degree in computer technology with the School of Computer Science, Hubei University of Technology, Wuhan.

His research interests include intelligent remote sensing image processing, image fusion, and deep learning.

Lu Wei received the B.Eng. degree in software engineering from the Wuhan University of Technology, Wuhan, China, in 2006, and the M.Eng. degree in computer technology from Wuhan University, Wuhan, in 2019.

She was a Senior Engineer with Huawei Technologies Corporation and is currently an Associate Professor with the School of Information Science and Engineering, Wuchang Shouyi University, Wuhan. Her research interests include image radiance correction, multisensor image fusion, image quality assessment, and intelligent image processing.

Zhiqi Zhang received the B.Sc. degree in geographic information systems from Huazhong Agricultural University, Wuhan, China, in 2006, the B.Eng. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, in 2006, and the M.Eng. degree in computer technology and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2015 and 2018, respectively.

He is currently an Associate Professor with the School of Computer Science, Hubei University of Technology, Wuhan. His research interests include system architecture, algorithm optimization, AI, and high-performance processing of remote sensing.

Xiaoxiao Feng received the B.Sc. degree in surveying and mapping from Southeast University, Nanjing, China, in 2014, the M.Sc. degree in earth exploration and information technology from the China University of Geology, Wuhan, China, in 2017, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2021.

She is currently a Lecturer with the School of Computer Science, Hubei University of Technology, Wuhan, China. Her research interests include high spatial resolution and hyperspectral remote sensing image processing and analysis.

Shao Xiang received the B.S. degree in automation engineering from Hefei University, Hefei, China, in 2017, and the M.S. degree in automation from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2020. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

His research interests include change detection, image compression, image fusion, object detection, and semantic segmentation of remote sensing.