Cross-Scale Interaction with Spatial-Spectral Enhanced Window Attention for Pansharpening

Hangyuan Lu, Huimin Guo, Rixian Liu, lingrong Xu, Weiguo Wan, Wei Tu, Yong Yang, *Senior Member, IEEE*

***Abstract*—Pansharpening is a process that fuses a multispectral (MS) image with a panchromatic (PAN) image to generate a high-resolution multispectral (HRMS) image. Current methods often overlook scale inconsistency and the correlation within and between a window domain, resulting in suboptimal outcomes. Additionally, the use of deep CNN or Transformer often leads to high computational expenses. To address these challenges, we present a lightweight pansharpening network that leverages cross-scale interaction and spatial-spectral enhanced window attention. We first design a spatial-spectral enhanced window Transformer (SEWformer) to effectively capture crucial attention within and between interleaved windows. To improve scale consistency, we develop a cross-scale interactive encoder that interacts with different scale attentions derived from the SEWformer. Furthermore, a multiscale residual network with channel attention is constructed as a decoder, which, in conjunction with the encoder, ensures precise detail extraction. The final HRMS image is obtained by combining the extracted details with the upsampled MS image. Extensive experimental validation on diverse datasets showcases the superiority of our approach over state-of-the-art pansharpening techniques in terms of both performance and efficiency. Compared to the second-best comparison approach, our method achieves significant improvements in the ERGAS metric: 29.6% on IKONOS, 43.8% on Pléiades, and 27.6% on WorldView-3 datasets. The code for this work is available at https://github.com/yotick.**

***Index Terms*—Pansharpening, cross scale, self-attention, Transformer.**

I. Introduction

DUE to limitations in payload capacity and bandwidth, satellite sensors face challenges in capturing images with both high spectral and spatial resolutions simultaneously [1]. Satellites like IKONOS, WorldView-2, and WorldView-3 acquire low-resolution multispectral (MS) images and high-resolution panchromatic (PAN) images separately. However, utilizing only MS or PAN images independently for applications leads to the loss of valuable information. Hence, the pansharpening, which seeks to combine a PAN image and an MS image to generate a high-resolution multispectral image (HRMS), has emerged as a crucial technique in the fields of remote sensing image processing. By integrating complementary information from PAN and LRMS images, the resulting HRMS image exhibits improved details and enhanced interpretability. This technique enables more accurate object identification, classification, and analysis in various domains such as urban planning, environmental monitoring, agriculture, and disaster management [2], [3].

Pansharpening approaches can be categorized into four classes: component substitution (CS)-based, multi-resolution analysis (MRA)-based, variational optimization (VO)-based, and deep learning (DL)-based methods [4]. CS-based approaches involve projecting the original multispectral (MS) image into another domain and replacing its spatial component with the PAN image. Representative CS methods include Gram Schmidt Adaptive (GSA) [5], robust band-dependent spatial detail (RBDSD) [6], and intensity-hue-saturation transform (IHS) [7]. However, CS-based methods are prone to produce spectral distortion [8]. MRA-based methods utilize a multi-scale decomposition of the PAN image to extract its spatial components. Recent advancements include a revised version of additive wavelet luminance proportional (AWLP-R) [9], regression based Laplacian pyramid [10], and fuzzy logic-based guided filter decomposition [11], which can improve spectral quality to a certain extend. Nevertheless, MRA-based methods are susceptible to spatial distortion [12], [13].

VO-based methods have gained popularity due to their ability to handle complex optimization problems and their flexibility in incorporating additional constraints and prior knowledge. The Bayesian probability model [14], sparse representation [15] and variational model [16] all belongs to this type of approach. Combining with detail injection model, Lu et al. [17] proposed an adaptive gradient and detail correction model to refine the spatial structure of the HRMS image, Xiao et al. [3] further presented a coefficient estimation model based on context-aware details injection fidelity (CDIF), which obtains impressive fusion result. These methods have demonstrated promising results with improved visual quality and enhanced spectral accuracy. However, the effectiveness of VO-based methods heavily relies on the suitability of the constructed model and the precise setting of parameters, which can be challenging to optimize [13], [16].

With the advancements in deep neural networks, DL-based pansharpening methods have demonstrated their ability to effectively capture complex relationships and extract meaningful features, thereby preserving both spectral and spatial

This work is supported by the National Natural Science Foundation of China (No.62362035, No.62261025, and No.62361030), and in part by the Project of the Education Department of Jiangxi Province (GJJ2201330). *(Hangyuan Lu and Rixian Liu contributed equally to this work.)(Corresponding author: Hangyuan Lu.)*

H. Lu, H. Guo, R. Liu, and L.Xu are with the College of Information Engineering, Jinhua Polytechnic, and also with Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua 321007, China (e-mail: lhyhziee@163.com; Guohm1995@163.com; liurixian@163.com; lynxu223@163.com).

Weiguo Wan is with School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang 330038, China(wanwgplus@163.com)

Wei Tu is with School of Big Data Science, Jiangxi Science and Technology Normal University, Nanchang 330038, China (ncsytuwei@163.com)

Yong Yang is with the School of Computer Science and Technology, Tiangong University, Tianjin 300387, China (e-mail: greatyangy@126.com).

information. For instance, Yang et al. [18] introduced PanNet, which surpasses traditional methods by integrating up-sampled multispectral images and training network parameters in the high-pass filtering domain. Ozcelik et al. [19] presented a self-supervised learning GAN framework called PCGAN, which treats pansharpening as a colorization task for PAN images. This approach demonstrates promising spatial quality. Tu et al. [20] utilized a clique structure-based multiscale dilated block and a multi-distillation residual information block, effectively integrating the information from both MS and PAN images. To enhance interpretability, Deng et al. [21] proposed detail injection-Based deep convolutional neural networks (DCNNs) which combined CNN with traditional fusion schemes to estimate nonlinear injection details, while Wu et al. [22] presented VO+Net, which incorporates spatial and spectral fidelity terms along with a weighted regularization term in DL network. DL-based methods generally outperform traditional approaches in pre-serving spectral fidelity, although they come with increased computational complexity and require a larger number of samples [23].

The Transformer architecture has gained significant attention and success in computer vision tasks. Its self-attention mechanism allows for modeling global spatial dependencies, making it promising for pansharpening. Meng et al. [24] proposed a purely Transformer-based model for pansharpening, utilizing self-attention to capture long-distance dependencies and achieve competitive performance compared to CNN-based methods. Hou et al. [25] proposed PAN-guided multiresolution fusion (PMRF) network which incorporates with Swin Transformer to exploit self-similarity and improve feature representation. While Transformer-based methods can outperform others on specific datasets, they require extremely high computational resources and rely heavily on large amounts of data [26].

Transformer-based pansharpening methods often rely on multiple Transformer blocks to enhance global attention. However, this approach introduces a trade-off by increasing the computational resources. Moreover, remote sensing images typically cover large areas, where pixels that are far apart exhibit weak correlations, while those within a window-area demonstrate stronger correlations. To extract more related attention, we propose a spatial-spectral enhanced window Transformer (SEWformer) by embedding extra spatial-spectral linear attention into the interleaved window Transformer. Further, due to the inconsistence in the fusion of different scale, a cross scale interactive encoder with SEWformer is presented to narrow the gap between the fusion of different scales. In addition, a decoder combining multiscale residual network with channel attention is designed to extract accurate details. The contributions of this work are outlined as follows:

1) A lightweight network for pansharpening proposed by leveraging cross-scale interaction and spatial-spectral enhanced window attention, which can improve the performance and efficiency in both reduced- and full-scale image fusion.

2) We present the SEWformer by incorporating extra spatial-spectral linear attention, which efficiently captures important attention within and between interleaved windows.

3) We propose a cross-scale interactive encoder that interacts with attentions derived from the SEWformers at different scales, which can narrow the gap between the fusion of different scales and enhances the overall fusion quality.

4) A decoder is designed by combining a multiscale residual network with channel attention to accurately extract image details, ensuring high-quality fusion results.

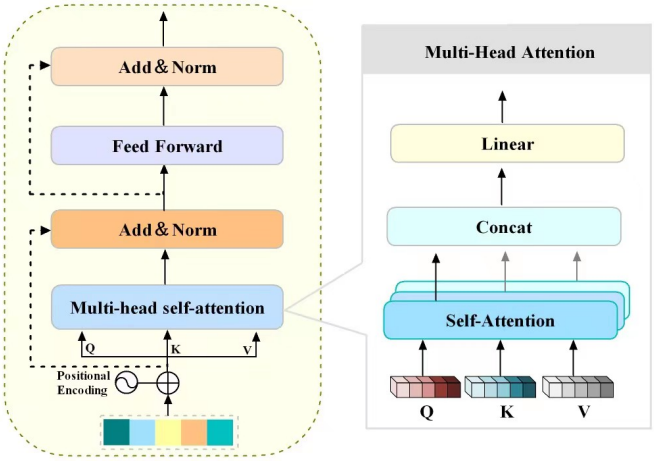

Fig. 1. Structure of Transformer

II. Related Work

The Vision Transformer (ViT), introduced by Dosovitskiy et al. applies the Transformer architecture to image recognition tasks. Unlike traditional convolutional neural networks, which rely on convolutional layers for feature extraction, ViT divides the input image into fixed-size patches and treats them as tokens. These patches are then processed by a sequence of Transformer layers, allowing the model to capture both global dependencies within the image.

A core component of ViT is the multi-head self-attention (MHSA) mechanism, as shown in Fig. 1. This mechanism enables the model to attend to different parts of the input sequence simultaneously. In the MHSA block with m heads, the input x is linearly projected into query (q), key (k), and value (v) representations using weight matrices W_q, W_k, and W_v respectively. This can be expressed as follows:

$$q = xW_q, k = xW_k, v = xW_v. \tag{1}$$

Next, for each head $i, i \in \{1, 2, ..., m\}$, the self-attention operation is performed by calculating the dot product of the query and key vectors scaled by the square root of the dimensionality (d). The SoftMax function is applied to normalize the dot product scores, resulting in attention weights (z_i) for each head. These attention weights are used to weight the corresponding value vectors. Finally, the attention-weighted value vectors are linearly combined and concatenated to obtain the multi-head attention (z) using the weight matrix W_o. The expression is as follows.

$$z_i = SoftMax\left(\frac{q_i.{k_i}^T}{\sqrt{d}}\right).v_i, \tag{2}$$

$$z = Cat(z_1, z_2, \cdots, z_m)W_o, \tag{3}$$

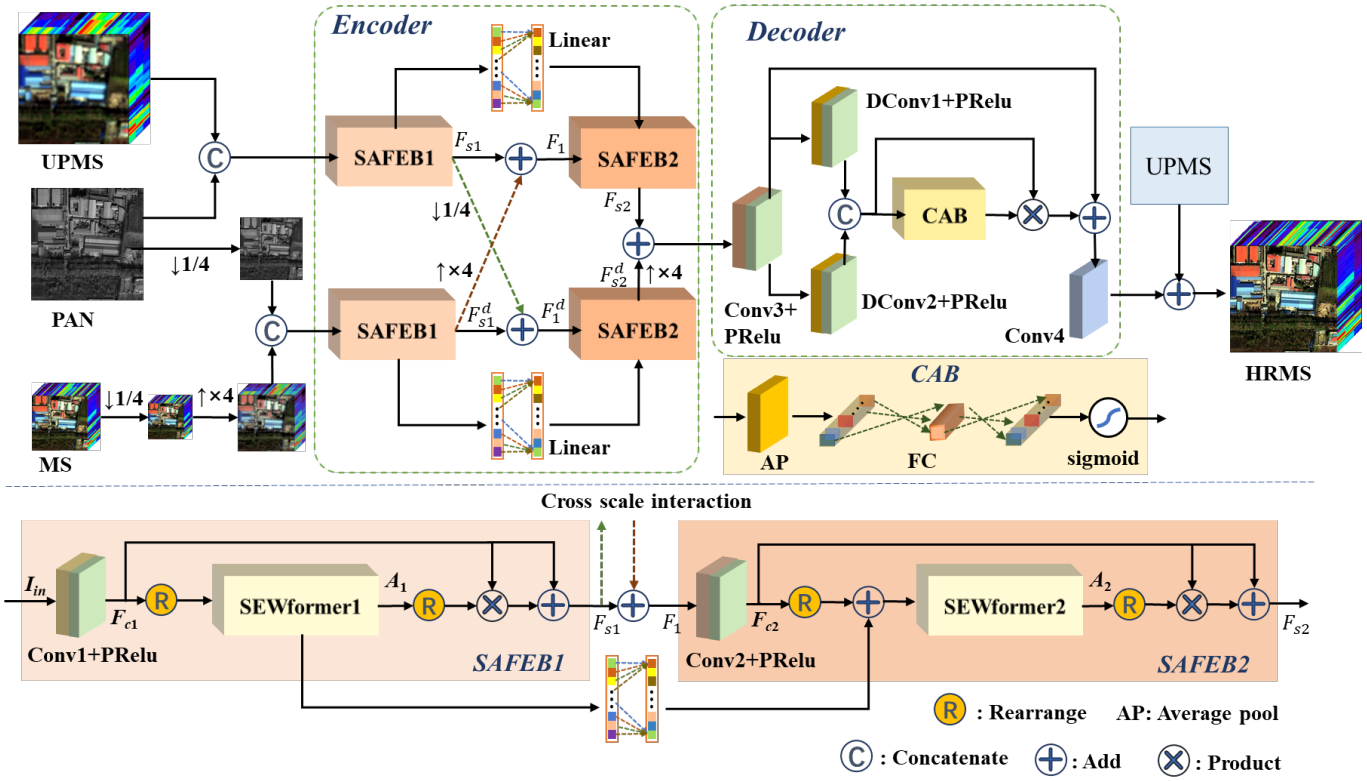

Fig. 2. Flowchart of the proposed method

where $Cat(\cdot)$ represents concatenation operation in channel dimension. The outputs of the multi-head attention module are passed through the add and norm layer and feed-forward layer to obtain the final attention representation, as shown in Fig. 1. ViT has achieved impressive results on various image recognition benchmarks and demonstrated its ability to model long-range dependencies effectively [26], [27]. However, the Transformer still presents challenges in applications due to its high memory usage. This serves as the motivation behind our work, where we aim to design an efficient Window-attention mechanism with good generalization and high accuracy.

III. Proposed Method

A. Overall Structure

The proposed model follows an encoder-decoder structure, as shown in Fig. 2. Majority existing pansharpening methods overlook the scale inconsistency between reduced-scale (RS) and full-scale (FS) image fusion, leading to unsatisfactory fusion results. To address this issue, the paper introduces an encoder with two feature extraction branches, each operating at different scales, to enhance cross-scale interaction. In the first branch, the input is the concatenation of the upsampled MS (UPMS) image and the PAN image. In the second branch, the input consists of the RS versions of the MS and PAN images, obtained using Wald's protocol [28]. Both branches primarily consist of two-stage self-attention feature extraction blocks (SAFEBs), namely SAFEB1 and SAFEB2. In the first stage, the interaction between different scales is facilitated by cross addition of the SAFEB1 outputs from the respective branches. This interaction can be expressed as follows.

$$\begin{cases} F_1 = F_{s1} + F_{s1}^d \uparrow \\ F_1^d = F_{s1}^d + F_{s1} \downarrow \end{cases}, \tag{4}$$

where F_{s1} and F_{s1}^d represent the outputs of SAFEB1 from original-scale and reduced-scale branches, respectively. F_1 and F_1^d represent the features after the interaction from the two branches, respectively. $\uparrow$ and $\downarrow$ denote up-sampling and down-sampling interpolation operations, respectively.

For the second stage, the outputs of the two branches are added together to further integrate the features at different scales. This integration can be expressed as follows.

$$O_{en} = F_{s2} + F_{s2}^d \uparrow, \tag{5}$$

where O_{en} represents the output of the encoder. F_{s2} and F_{s2}^d represent the features obtained by SAFEB2 from the two branches, respectively, as shown in Fig. 2.

The output of the encoder is then passed to the decoder, which comprises multiple residual networks with channel attention. The decoder aims to capture accurate details from the encoded features. The obtained details are subsequently integrated into the UPMS image, yielding the final HRMS image.

B. The Two-Stage SAFEBs

As illustrated in Fig. 2, the cascaded two-stage SAFEBs are designed as two residual SEWformer networks, working in tandem to effectively extract the self-attention features. In

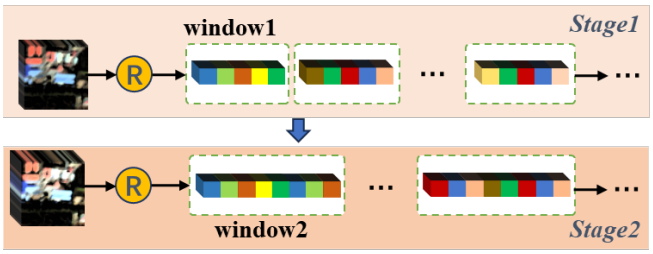

Fig. 3. Interleaved window mechanism.

the first stage, the input data undergoes a 24×3×3 convolution and PRelu layers, which replace the conventional position embedding operation. This substitution aims to enhance efficiency and improve the representation of local features. The resulting features, denoted as F_{c1} are then fed into SEWformer to compute window attention weights A_1. These weights are then rearranged to match the dimensions of F_{c1}. The rearranged attention is element-wise multiplied and added to F_{c1} to emphasize important spectral-spatial information and facilitate the learning of residual information. The resulting features F_1 can be defined as follows.

$$\begin{cases} F_{c1} = PR(conv(I_{in})) \\ A_1 = ATT(R(F_{c1})) \\ F_{s1} = F_{c1} + F_{c1} \times R(A_1) \end{cases}, \tag{6}$$

where I_{in} represents the input image, conv(·), PR(·), and R(·) represent convolution, PRelu, and rearrange operations, respectively. ATT(·) denotes the SEWformer process.

To further optimize the features obtained from the first stage, F_{s1} is fed into the second stage, specifically SAFEB2, after cross-scale interaction. SAFEB2 has a similar structure to SAFEB1. Importantly, the two stages are designed as a multiscale interleaved window mechanism. In SAFEB2, a 48×3×3 convolutional layer is first used for channel expansion. Then, the extracted features are rearranged and vectorized. Next, the window sizes of the two-stage SEWformers are set to different values, creating interleaved windows, as depicted in Fig. 3. This configuration allows for enhanced connections across the windows and effectively reduces window artifacts.

Besides, to enhance the connection between two-stage attentions, the output of SEWformer1 is added to the input of SEWformer2 through a linear layer. Then, similar to the process of obtaining F_{s1}, the features optimized by SEWformer F_{s2} can be defined as follows.

$$\begin{cases} F_{c2} = PR(conv(F_1)) \\ A_2 = ATT(R(F_{c2}) + LN(A_1)) \\ F_{s2} = F_1 + F_1 \times R(A_2) \end{cases}, \tag{7}$$

where LN(·) represents a linear layer.

C. SEWformer

Traditional Transformer-based models process input sequences in the global domain, resulting in high computational requirements and limited attention extraction in localized regions or windows. This paper addresses these limitations by presenting SEWformer, as illustrated in Fig. 4. In SEWformer,

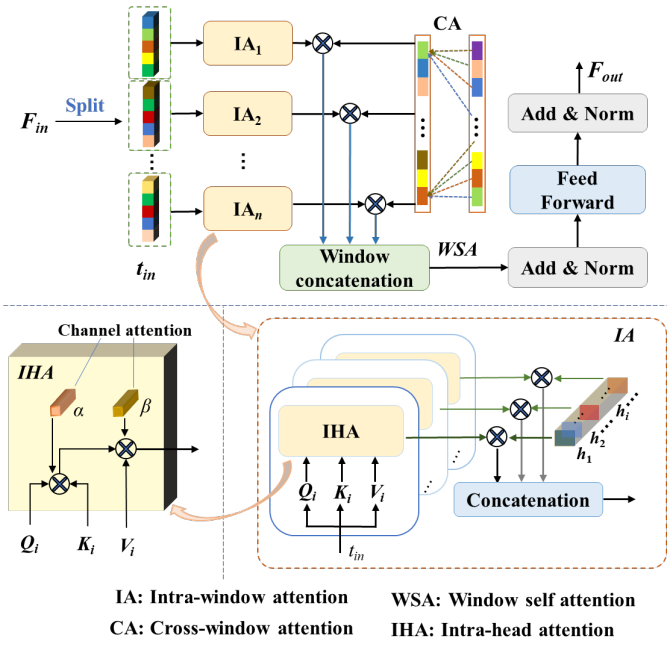

Fig. 4. Flowchart of the SEWformer

the input features F_{in} are initially vectorized into a sequence. This sequence is then divided into windows of a specific size, with each window containing a subset of tokens. Next, the tokens within each window undergo intra-window attention (IA) to enhance pixel dependencies within the window. However, since the windows are processed independently and lack communication, we propose a cross-window attention (CA) utilizing a linear layer to establish connections between the windows. By combining IA and CA, the resulting window self attention (WSA) is defined as follows:

$$WSA = Cats(IA(t_{in}) \times CA(\omega)), \tag{8}$$

where $Cats(\cdot)$ represents concatenation operation in spatial dimension. t_{in} and ω represent the input vectors of IA and CA, respectively. Similar to traditional transformers, WSA is further processed through normalization (Norm) and feedforward operations to obtain final attention map F_{out}.

The aforementioned CA can improve the spatial connections between windows. To further enhance spectral attention, we propose the inter-head attention and channel attention within each head. Specifically, we define learnable attention vectors for inter-head attention, which are multiplied with the multi-head attention outputs to enhance attention across heads. Within each head, we define learnable channel attention vectors α and β, which are respectively multiplied with K and V matrices to obtain enhanced intra-head attention (IHA). The process of IHA and IA can be defined as follows:

$$\begin{cases} IA(t_{in}) = Cat(IHA(Q_i, K_i, V_i) \times h_i) \\ IHA(Q_i, K_i, V_i) = soft\max(\frac{Q_i(K_i \times \alpha)^T}{\sqrt{d_k}})(V_i \times \beta) \end{cases}, \tag{9}$$

where h represents the attention vector across the heads. Denoting the output of $CA(\omega)$ as γ and substituting Equation

(9) into equation (8), we can achieve the parallelized form of WSA as follows:

$$WSA = Cats(soft\max(\frac{Q(K\times Cat(\alpha,h\times\gamma))^T}{\sqrt{d_k}}) \\ (V \times Cat(\beta, h \times \gamma))). \quad (10)$$

In Equation (10), both $Cat(\alpha, h \times \gamma)$ and $Cat(\beta, h \times \gamma)$ are learnable 3-dimensional matrices. To simplify the notation, we denote these matrices as M_K and M_V, respectively, by concatenating their dimensions. Thus, Equation (10) can be rewritten as follows:

$$\begin{cases} WSA = Cats(soft\max(\frac{Q(K\times M_K)^T}{\sqrt{d_k}})(V \times M_V)) \\ M_K = Cat(\alpha, h, \gamma),\ M_V = Cat(\beta, h, \gamma) \end{cases}. \quad (11)$$

Based on Equation (11), both the attention matrices M_K and M_V that contain the spatial attention cross the window and the channel attention in the window can be optimized simultaneously during the training process, thereby further enhance the window-wise attention.

D. Decoder

To leverage the advantages of multiscale information and residue learning in enhancing the reconstruction of HRMS images, the decoder is designed following a multiscale residual network structure. The initial layer of the decoder consists of a 3×3 convolution and PRelu layer aimed at extracting the initial decoder representations. Subsequently, these initial representations are passed through two dilation convolutions, DConv1 and DConv2, with dilation rates of 2 and 3, respectively. This choice of dilation rates allows the convolutions to efficiently capture features with multiple receptive fields, enabling the network to incorporate information from a wider context. To incorporate multiscale information, the feature maps obtained from the dilation convolution layers are concatenated along the channel dimension. Additionally, in order to highlight important features at different scales, the concatenated feature undergoes a channel attention block (CAB) consisting of an average pooling (AP) layer, squeeze-then-extension FC layer, and sigmoid function. The process of CAB can be expressed as:

$$CA(F_D) = \sigma(FC_2(FC_1(AP(F_D)))), \quad (12)$$

where F_D represents the input decoder feature, $\sigma(\cdot)$ represents a sigmoid function. Furthermore, a residual connection is established by element-wise addition of the channel attention feature with the original features. This residual connection allows the decoder to focus on learning the difference between the input and the target images. The feature maps obtained from the residual structure are then passed through a final convolution layer, generating the reconstructed detail information.

The HRMS image is obtained by incorporating the detail information into the UPMS image. The L1 loss function is chosen due to its robustness to outliers in the training process.

IV. Experimental Analysis

A. Experimental Setting

We conducted a comprehensive evaluation of the model's effectiveness using the Pléiades, IKONOS, and WorldView-3 datasets, as outlined in Table I. The evaluation encompassed both RS and FS experiments. In the RS experiment, we applied filters specific to each sensor's modulation transfer function to degrade the source images. These degraded images were subsequently downsampled by a factor of four. As outlined in Wald's protocol [28], the original MS images served as the ground truth (GT). In the FS experiment, we performed image fusion at the original scale since no GT image was available.

TABLE I
Details of Datasets

Sensor	Pléiades	IKONOS	WorldView-3
MS/PAN resolutions	0.5/2.0(m)	0.82/3.2(m)	0.31/1.24(m)
MS sizes (RS/FS)	64×64×4/ 256×256×4	64×64×4/ 256×256×4	64×64×8/ 256×256×8
PAN sizes (RS/FS)	256×256/ 1024×1024	256×256/ 1024×1024	256×256/ 1024×1024
MS bands	red(R), green(G), blue(B), near infrared(NIR)	R, G, B, NIR	R, G, B, NIR1, NIR2, coastal blue red edge, yellow

To train our proposed model, we assembled a dataset comprising 2500 groups of RS images from each dataset. Each group consisted of degraded PAN and multispectral MS images, along with their corresponding GT references. To augment the training dataset, we employed random cropping, extracting blocks of size 64×64, which was half the original size. Considering GPU performance, we selected a batch size of 25. The training process involved 300 epochs, commencing with a learning rate of 0.0005. To ensure stable convergence, the learning rate was halved every 100 epochs. Additionally, we curated an independent test set consisting of 100 image pairs from each dataset for evaluation purposes.

To assess the effectiveness of our method, we conducted a comparative analysis with various state-of-the-art (SOTA) techniques, including both traditional and deep learning-based approaches. The traditional methods considered in the comparison were CDIF [3], RBDSD [6], and AWLP-R [9]. The deep learning-based methods included APNN-FT [30], VO+Net [22], FusionNet [21], PCGAN [19], TDNet [31], and PMRF [25]. Specifically, PMRF is a recently published method based on Swin Transformer for a more comprehensive evaluation. To ensure a fair comparison, we retrained the deep learning-based methods using our datasets. For spectral benchmarking, we utilized the results obtained by the EXP method [29] as a reference.

The comparisons were conducted on a computer equipped with an RTX-3060 GPU and 24 GB of RAM. The evaluation metrics used for the RS experiments included UIQI↑, Q2n↑, ERGAS↓, SAM↓, and SCC↑ [2], [32]. In the FS experiments, the evaluation metrics used were D_λ↓, D_s↓, and QNR↑ [33]. A higher value of the metric indicated a better outcome, except for the metrics with ↓, where a lower value indicated a superior result.

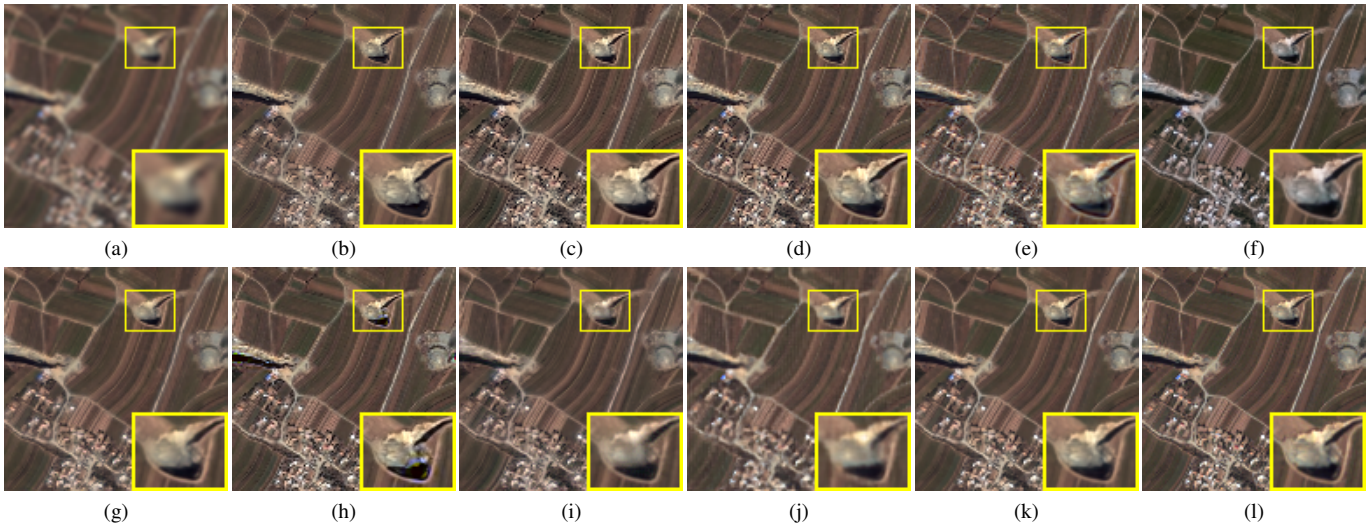

Fig. 5. Pansharpened RS images in IKONOS dataset. (a) EXP. (b) AWLP-R. (c) RBDSD. (d) CDIF. (e) APNN-FT. (f) PCGAN. (g) FusionNet. (h) VO + Net. (i) TDNet. (j) PMRF. (k) Proposed. (l) GT.

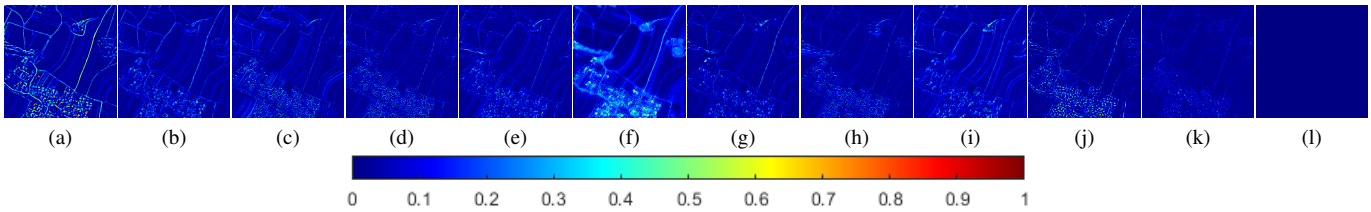

Fig. 6. AEMs corresponding to the images depicted in Fig. 5. (a) EXP. (b) AWLP-R. (c) RBDSD. (d) CDIF. (e) APNN-FT. (f) PCGAN. (g) FusionNet. (h) VO + Net. (i) TDNet. (j) PMRF. (k) Proposed. (l) GT.

TABLE II
OBJECTIVE ASSESSMENT OF FUSION RESULTS PRESENTED IN FIG. 5 AND MEAN ASSESSMENT ON IKONOS DATASET

Methods	Assessment of Fig. 5					Mean Assessment				
	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑
EXP [29]	0.7475	3.3952	4.5409	0.6708	0.7518	0.7627	3.3579	4.0487	0.6849	0.7574
AWLP-R [9]	0.8898	3.6511	3.0449	0.8911	0.8981	0.8682	3.5731	3.3621	0.8540	0.8633
RBDSD [6]	0.8948	4.1351	3.2610	0.8955	0.8901	0.8529	4.4319	3.7717	0.8410	0.8427
CDIF [3]	0.9294	3.0517	2.7942	0.9067	0.9278	0.9068	3.3128	2.8697	0.8760	0.9033
APPN-FT [30]	0.9047	4.0734	3.0329	0.8754	0.9035	0.8780	4.2710	3.3817	0.8449	0.8673
PCGAN [19]	0.9202	7.1009	3.7293	0.9506	0.9207	0.8412	8.4360	6.0984	0.9263	0.8224
FusionNet [21]	0.9325	3.0577	2.4453	0.9242	0.9348	0.9118	3.1394	2.6173	0.8973	0.9088
VO+Net [22]	0.9336	3.4027	2.7799	0.9100	0.9314	0.8944	3.5710	3.1333	0.8685	0.8866
TDNet [31]	0.8760	4.4087	3.0700	0.9080	0.8819	0.8877	3.7885	3.0256	0.8896	0.8783
PMRF [25]	0.8818	3.2283	3.6040	0.7812	0.8699	0.8843	3.2154	3.0925	0.8252	0.8680
Proposed	**0.9632**	**2.4309**	**1.8080**	**0.9540**	**0.9637**	**0.9480**	**2.4615**	**1.8424**	**0.9380**	**0.9460**

B. Reduced-scale Experiments

Fig. 5 presents the pansharpened results for a test sample from the IKONOS dataset, showcasing solely the RGB bands to enhance visual impact. The enlarged yellow rectangles provide close-ups of smaller regions. A visual inspection reveals that our method achieves superior visual quality. Specifically, the result obtained from the PCGAN method exhibits noticeable spectral distortion, which in turn introduces a significant bias in the overall color representation. The fusion results of FusionNet, TDNet, and PMRF appear blurred, indicating insufficient extraction of spatial information and resulting in spatial distortion. Moreover, the results of AWLRP-R, RBDSD, CDIF, and VO+Net exhibit some additional artifacts. Notably, the result of proposed method is visually closer to the GT.

Fig. 6 depicts the Absolute Error Maps (AEMs) of the compared methods, which offer a clearer visualization of the discrepancies in the results obtained by the different methods. It is evident that the proposed method exhibits fewer artifacts, and the fused result of the proposed method bears a closer resemblance to the GT.

Table II provide a quantitative assessment of the fusion results from IKONOS dataset. The table is divided into two sections: the left section corresponds to the fusion results displayed in Fig. 6, and the right section presents the mean objective assessment from all 100 testing samples. The second-

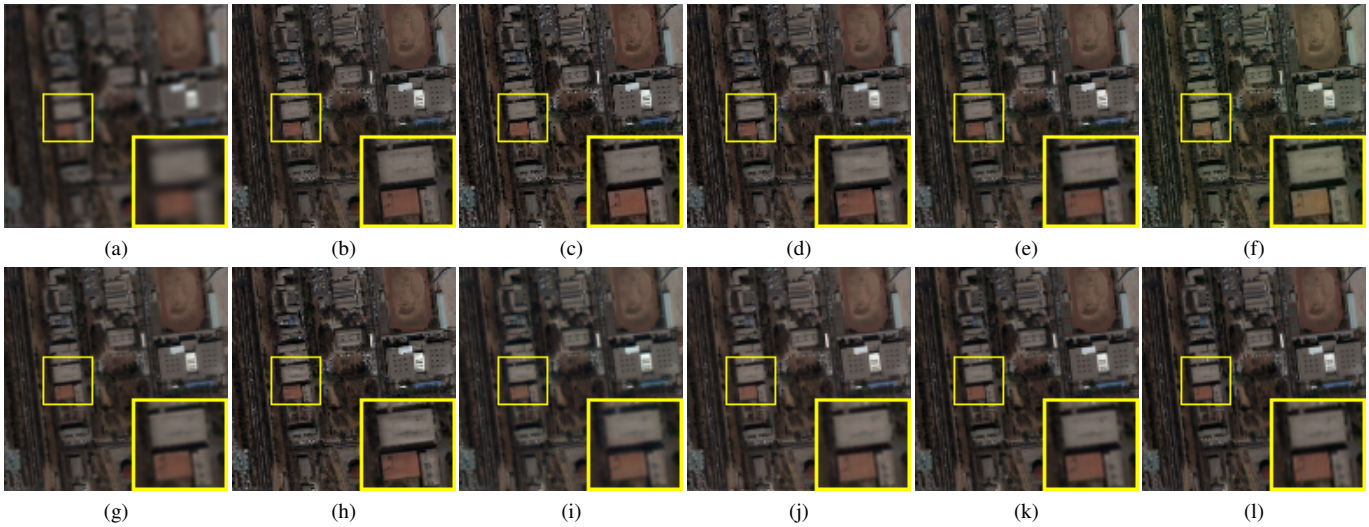

Fig. 7. Pansharpened RS images in Pléiades dataset. (a) EXP. (b) AWLP-R. (c) RBDSD. (d) CDIF. (e) APNN-FT. (f) PCGAN. (g) FusionNet. (h) VO + Net. (i) TDNet. (j) PMRF. (k) Proposed. (l) GT.

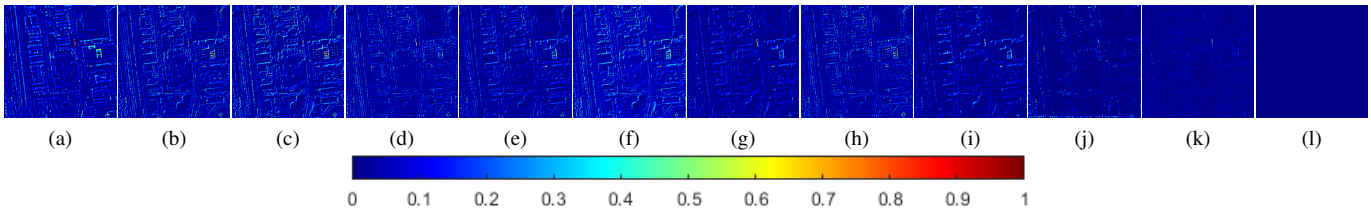

Fig. 8. AEMs corresponding to the images depicted in Fig. 7. (a) EXP. (b) AWLP-R. (c) RBDSD. (d) CDIF. (e) APNN-FT. (f) PCGAN. (g) FusionNet. (h) VO + Net. (i) TDNet. (j) PMRF. (k) Proposed. (l) GT.

TABLE III
OBJECTIVE ASSESSMENT OF FUSION RESULTS PRESENTED IN FIG. 7 AND MEAN ASSESSMENT ON PLÉIADES DATASET

Methods	Assessment of Fig. 7					Mean Assessment				
	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑
EXP [29]	0.8203	3.3301	5.1453	0.7538	0.8246	0.8194	3.0838	3.9728	0.7522	0.8173
AWLP-R [9]	0.8452	3.7380	5.7099	0.7700	0.8380	0.8863	3.0752	3.4263	0.8332	0.8846
RBDSD [6]	0.8216	4.8829	6.4371	0.7624	0.8137	0.8774	3.5755	3.6265	0.8354	0.8730
CDIF [3]	0.8878	3.5868	4.7984	0.8240	0.8815	0.9150	2.7529	2.8492	0.8620	0.9133
APPN-FT [30]	0.8873	4.5985	4.9075	0.8181	0.8831	0.9055	3.6274	3.0794	0.8533	0.9014
PCGAN [19]	0.8302	10.1973	6.8884	0.7783	0.8261	0.8653	7.7779	5.0045	0.8483	0.8052
FusionNet [21]	0.9073	3.3756	4.1828	0.8632	0.9046	0.9282	2.8778	2.9819	0.9039	0.9262
VO+Net [22]	0.8884	3.8272	5.0128	0.8091	0.8813	0.9122	2.7449	2.9123	0.8660	0.9090
TDNet [31]	0.8996	4.4728	4.5817	0.8318	0.8899	0.9215	3.5946	2.6642	0.9095	0.9164
PMRF [25]	0.9641	3.2155	3.3897	0.9448	0.9413	0.9508	2.8368	2.6512	0.9451	0.9122
Proposed	**0.9843**	**2.7567**	**1.7214**	**0.9734**	**0.9836**	**0.9628**	**2.3943**	**1.4894**	**0.9616**	**0.9629**

best outcomes are emphasized in blue, while the optimal values are marked in bold. Notably, the proposed method exhibits pronounced advantages across all metrics. The proposed model's effectiveness is reinforced by the objective evaluation.

The fusion results of a test sample from the Pléiades dataset are visualized in Fig. 7. It is evident that the result obtained from the PCGAN method exhibits significant spectral distortion throughout the image. The results of RBDSD and AWLP-R display oversaturated colors on the orange rooftops. CDIF and VO+Net introduce additional black artifacts to the gray rooftops. The results of TDNet and FusionNet appear somewhat blurry, while that of PMRF seems overbright. In contrast, the proposed method closely resembles the GT. This observation is further supported by the residual maps depicted in Fig. 8. Overall, the proposed method strikes the best balance between spatial enhancement and spectral preservation. The quantitative evaluations, as shown in Table III, clearly indicate that the proposed method outperforms all other methods in terms of all metrics, boasting a considerable advantage.

The pansharpened results of the comparison methods on the WorldView-3 dataset are depicted in Fig. 9 . Upon closer observation of the enlarged yellow rectangles, it is clear that PCGAN introduces extra artifacts especially on the roofs. The results obtained with APNN-FT, FusionNet, TDNet, and PMRF exhibit slight blurriness and some degree of spatial distortion. In comparison to the GT, the results of AWLPR,

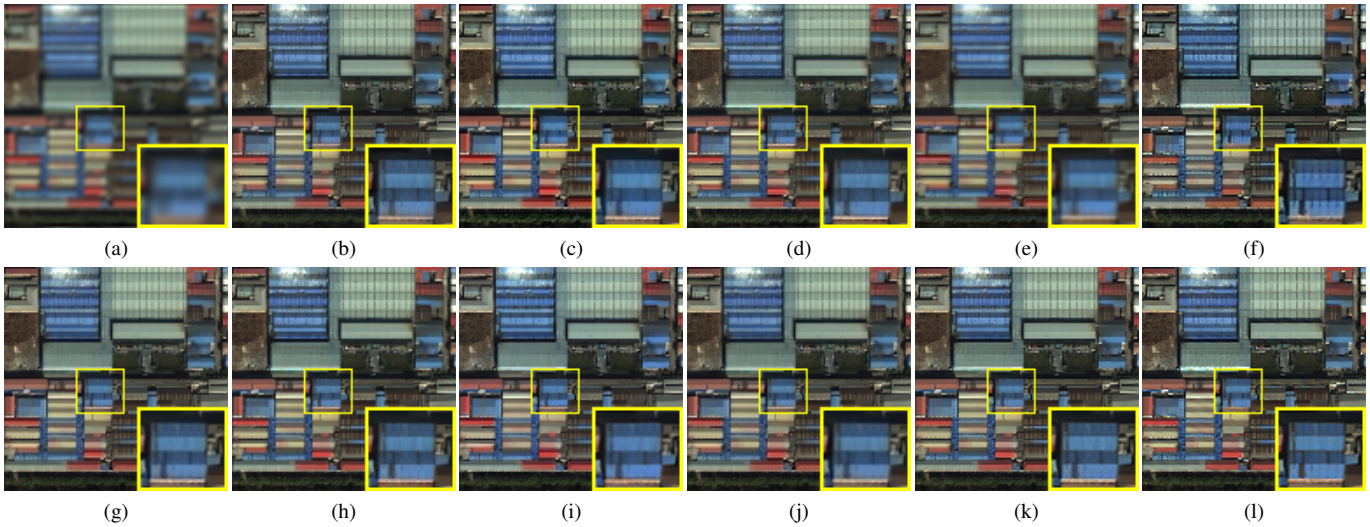

Fig. 9. Pansharpened RS images in WorldView-3 dataset. (a) EXP. (b) AWLP-R. (c) RBDSD. (d) CDIF. (e) APNN-FT. (f) PCGAN. (g) FusionNet. (h) VO + Net. (i) TDNet. (j) PMRF. (k) Proposed. (l) GT.

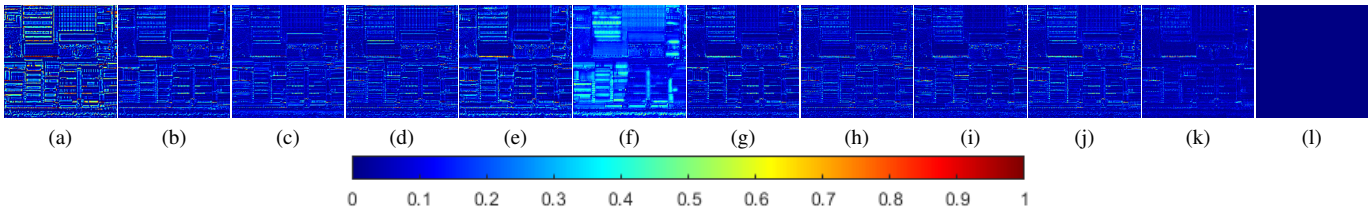

Fig. 10. AEMs corresponding to the images depicted in Fig. 9. (a) EXP. (b) AWLP-R. (c) RBDSD. (d) CDIF. (e) APNN-FT. (f) PCGAN. (g) FusionNet. (h) VO + Net. (i) TDNet. (j) PMRF. (k) Proposed. (l) GT.

TABLE IV
OBJECTIVE ASSESSMENT OF FUSION RESULTS PRESENTED IN FIG. 9 AND MEAN ASSESSMENT ON WORLDVIEW-3 DATASET

Methods	Assessment of Fig. 9					Mean Assessment				
	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q8↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q8↑
EXP [29]	0.7048	5.7806	6.7640	0.6826	0.7205	0.6567	5.4430	6.1138	0.6284	0.6553
AWLP-R [9]	0.9141	5.8583	4.0613	0.8924	0.9186	0.8769	5.4589	4.0417	0.8599	0.8805
RBDSD [6]	0.9184	6.7106	4.1083	0.8901	0.9245	0.8731	6.2859	4.1564	0.8539	0.8796
CDIF [3]	0.9206	5.6731	3.8896	0.8921	0.9241	0.8875	5.1814	3.7217	0.8641	0.8914
APPN-FT [30]	0.8863	6.9949	4.6100	0.8854	0.8940	0.8065	6.7477	4.7598	0.8091	0.8208
PCGAN [19]	0.9255	6.9740	2.8755	0.9415	0.9214	0.8524	8.3228	5.6929	0.8971	0.8302
FusionNet [21]	0.9210	5.9649	3.9066	0.8904	0.9279	0.8903	5.1457	3.7813	0.8736	0.8936
VO+Net [22]	0.9336	5.7136	3.6217	0.9025	0.9367	0.8842	5.7228	3.8555	0.8657	0.8846
TDNet [31]	0.9319	5.9780	3.7485	0.8983	0.9359	0.8912	5.3467	3.7361	0.8787	0.8936
PMRF [25]	0.9233	6.0508	3.8426	0.8989	0.9309	0.8968	5.1627	3.6081	0.8797	0.9005
Proposed	**0.9639**	**5.1103**	**2.7428**	**0.9559**	**0.9657**	**0.9367**	**4.3248**	**2.6137**	**0.9445**	**0.9393**

RBDSD, CDIF, and VO+Net appear somewhat darker. Conversely, the proposed method yields the most favorable results. Fig. 10 further corroborates the superiority of our method in terms of residual aspects. It displays the AEMs of all fusion results in Fig. 9, and it's clear that our method contains fewer residual artifacts than the other methods.

The objective assessment of the WorldView-3 dataset is presented in Table IV. Similar to the results obtained from the IKONOS and Pléiades datasets, our method clearly outperforms in all metrics. These findings demonstrate that our approach consistently delivers superior results.

C. Full-scale Experiments

To validate the efficacy of our model on FS image fusion, we take the IKONOS dataset as an example and provide visual comparisons in Fig. 11. The fused image size is B×1024×1024, with B indicating the number of bands. Due to its large size, it is not feasible to directly display it in the paper. To enhance visualization, we focus on a specific area indicated by a yellow rectangle and provide a close-up view. Due to the absent of a GT image for the full resolution case, we rely on the UPMS method as a spectral benchmark.

Notably, the outcomes obtained from PCGAN, FusionNet, and TDNet exhibit significant spectral distortion, leading to noticeable changes in the color of the roofs. The results of

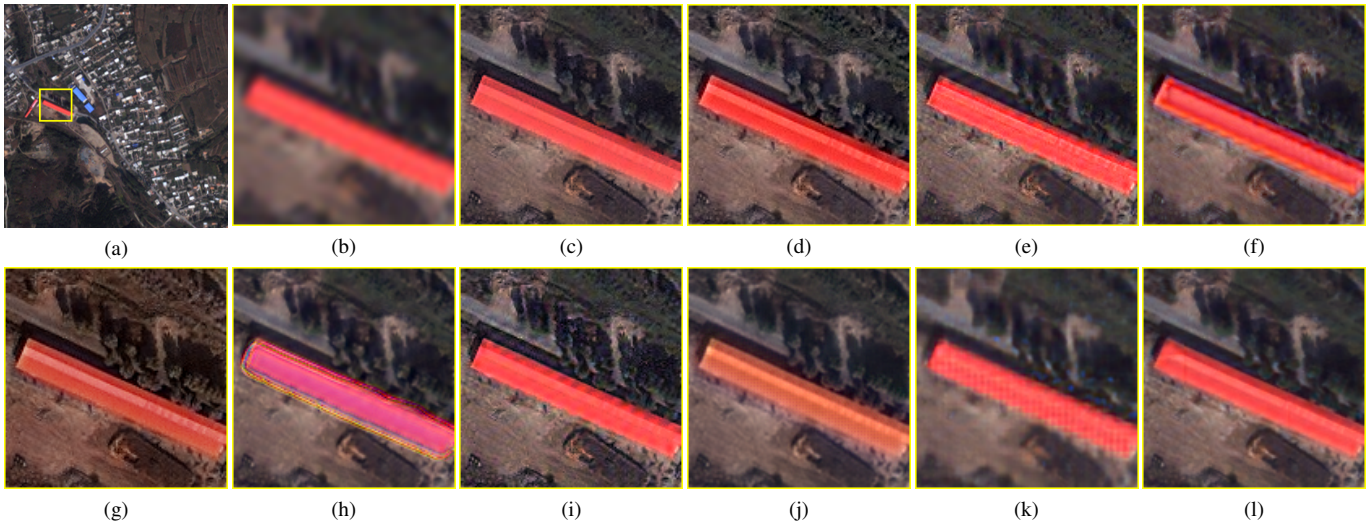

Fig. 11. Pansharpened FS images in IKONOS dataset. (a) FS image by EXP. (b) EXP. (c) AWLP-R. (d) RBDSD. (e) CDIF. (f) APNN-FT. (g) PCGAN. (h) FusionNet. (i) VO + Net. (j) TDNet. (k) PMRF (l) Proposed.

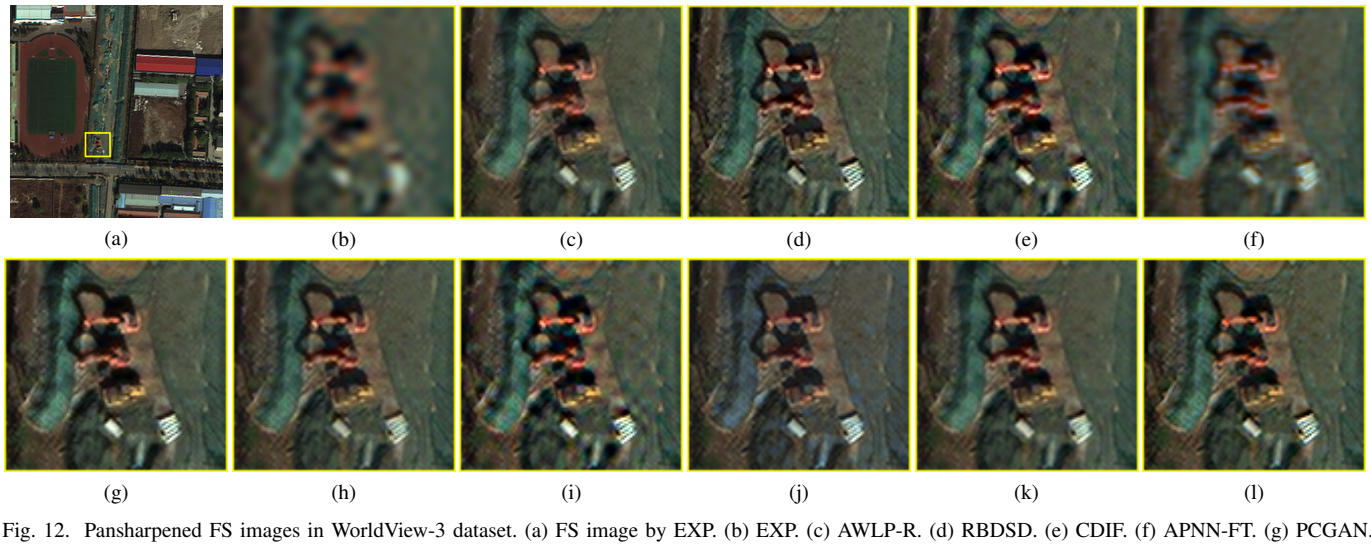

Fig. 12. Pansharpened FS images in WorldView-3 dataset. (a) FS image by EXP. (b) EXP. (c) AWLP-R. (d) RBDSD. (e) CDIF. (f) APNN-FT. (g) PCGAN. (h) FusionNet. (i) VO + Net. (j) TDNet. (k) PMRF (l) Proposed.

CDIF, APNN-FT, and VO+Net suffer from spatial distortion. The result of RBDSD appears over-saturated, while that of PMRF is blurred. In comparison, our method produces a more natural result and maintains better spectral fidelity, as shown in the comparison with the EXP reference.

To further quantify the fusion quality, we provide objective evaluation metrics in Table V. It can be observed that our method achieves the highest performance in terms of $D_s\downarrow$ and QNR$\uparrow$ metrics with a clear advantage, and ranks as the second-best in terms of the $D_\lambda\downarrow$ metric.

We also conducted the FS experiments on the WorldView-3 dataset, with the subjective results shown in Fig. 12. From the figure, we can observe that the result of APPN-FT appears blurred. The results of CDIF and VO+Net appear unnatural and suffer from over-sharpening in certain areas, leading to spectral distortion. The result of TDNet exhibits noticeable spectral distortion, as the overall color is visibly altered. In contrast, the result of our proposed method appears more natural, preserving better spectral and texture information compared to the other approaches. We further evaluate the performance through the objective assessment presented in Table VI. The results show that our method achieves the best performance in the D_λ and QNR metrics, and the second-best result in the D_s metric, in terms of mean assessment. These quantitative findings demonstrate that the proposed method enhances the fusion quality for both RS and FS images, affirming its effectiveness across different scales.

D. Ablation Study

In this section, we conducted an ablation study to assess the impact of the proposed components in our method. Our approach primarily incorporates SAFEB based on SEWformer and cross-scale interaction (CI) in the encoder, as well as multiscale dilation convolution (MDC) and CAB in the decoder. We evaluated the performance of several models: the base model with a single-branch convolutional encoder and decoder network (Base), the model with one-stage SAFEB (1SAFEB), the model with two-stage SAFEBs (2SAFEBs), the

TABLE V
QUANTITATIVE ASSESSMENT OF FS EXPERIMENTS ON THE IKONOS DATASET.

Methods	Assessment of Fig. 11			Mean Assessment		
	$D_\lambda\downarrow$	$D_s\downarrow$	QNR$\uparrow$	$D_\lambda\downarrow$	$D_s\downarrow$	QNR$\uparrow$
EXP [29]	0.0005	0.2254	0.7743	0.0006	0.2334	0.7662
AWLP-R [9]	0.1536	0.1399	0.7280	0.1456	0.1597	0.7195
RBDSD [6]	0.1647	0.1755	0.6888	0.1293	0.1385	0.7530
CDIF [3]	0.0970	0.0661	0.8433	0.0866	0.0653	0.8549
APPN-FT [30]	0.0693	0.0378	0.8943	0.0624	0.0639	0.8784
PCGAN [19]	0.1411	0.1966	0.6901	0.1305	0.2374	0.6676
FusionNet [21]	0.0894	0.0675	0.8491	0.0772	0.0663	0.8618
VO+Net [22]	0.1383	0.1162	0.7615	0.1246	0.1219	0.7713
TDNet [31]	0.1190	0.0778	0.8125	0.1044	0.0772	0.8267
PMRF [25]	**0.0131**	0.1461	0.8427	**0.0236**	0.1567	0.8233
Proposed	0.0642	**0.0299**	**0.9078**	0.0600	**0.0458**	**0.8971**

TABLE VI
QUANTITATIVE ASSESSMENT OF FS EXPERIMENTS ON THE WORLDVIEW-3 DATASET.

Methods	Assessment of Fig. 12			Mean Assessment		
	$D_\lambda\downarrow$	$D_s\downarrow$	QNR$\uparrow$	$D_\lambda\downarrow$	$D_s\downarrow$	QNR$\uparrow$
EXP [29]	0.0006	0.1159	0.8835	0.0004	0.0944	0.9053
AWLP-R [9]	0.1144	0.1271	0.7731	0.0821	0.0886	0.8376
RBDSD [6]	0.1167	0.1328	0.7660	0.0641	0.0999	0.8436
CDIF [3]	0.0479	0.0537	0.9009	0.0337	0.0367	0.9314
APPN-FT [30]	0.0737	**0.0371**	0.8919	0.0599	0.0439	0.8990
PCGAN [19]	0.1325	0.1320	0.7530	0.1008	0.1011	0.8094
FusionNet [21]	0.0494	0.1110	0.8451	0.0327	0.0723	0.8979
VO+Net [22]	0.0476	0.0375	0.9106	0.0351	**0.0303**	0.9359
TDNet [31]	0.0934	0.1308	0.7880	0.0660	0.0870	0.8534
PMRF [25]	0.0675	0.1089	0.8310	0.0398	0.0720	0.8918
Proposed	**0.0397**	0.0472	**0.9150**	**0.0282**	0.0315	**0.9395**

model with a two-branch CI encoder (2B_2SA_CI), the model with MDC added to 2B_2SA_CI (2B_2SA_CI_MDC), and the model with CAB added to 2B_2SA_CI_MDC (Proposed). By retraining these models and comparing them with the proposed model, we gain insights into the individual contributions of these components.

The subjective evaluation and corresponding AEMs are presented in Fig. 13. While it may be challenging to discern differences from the RGB images in Fig. 13, we primarily refer to AEMs with orange rectangles that highlight the variations in residue levels. The mean objective evaluations are summarized in Table VII. Notably, the objective evaluation aligns with the subjective evaluation results. The base model performs poorly, exhibiting the highest residue content. However, the addition of SAFEB significantly improves the model's performance. As the network depth increases, such as in the case of the 2SAFEBs models, the performance further improves. The introduction of the two-branch CI also leads to significant performance enhancement. Moreover, the design of CAB or MDC in the decoder contributes to improved performance. The proposed model, which integrates all components, achieves the highest performance across all metrics. The ablation study provides further confirmation of the necessity and effectiveness of the proposed components.

To validate the exchange of feature maps between different scales, and confirm the complementary nature of the multi-scale information, we further conducted the visualization experiments. The proposed method employs two-stage SAFEBs to facilitate the interaction between features at different scales. Taking the images shown in Fig. 13 as an exmple, the visualization of the 5th channel are shown in Fig. 14. F_{s1} and F_1 represent the outputs of the 1st-stage SAFEB in the original scale, before and after the scale-interaction, respectively. F_{s1}^d and F_1^d are the corresponding outputs in the degraded scale. F_{s2} is the output of the 2nd-stage SAFEB in the original scale, and $F_{s2}^d\uparrow$ is the upsampled output from the degraded scale. F_c is the combination of F_{s2} and $F_{s2}^d\uparrow$.

By interacting these multi-scale feature maps through the SAFEB modules, the network is able to integrate information across different scales. This is further evidenced in Fig. 14, where the features after interaction become more distinctive, and the fused features F_c contain a richer mixture of fine details and high-level semantics, suggesting a synergistic exchange of information across scales. The quantitative evaluation of the models with and without cross-scale interaction, shown in Table VII, also verifies the effectiveness of this interaction.

To further assess the effectiveness of the proposed SEW-

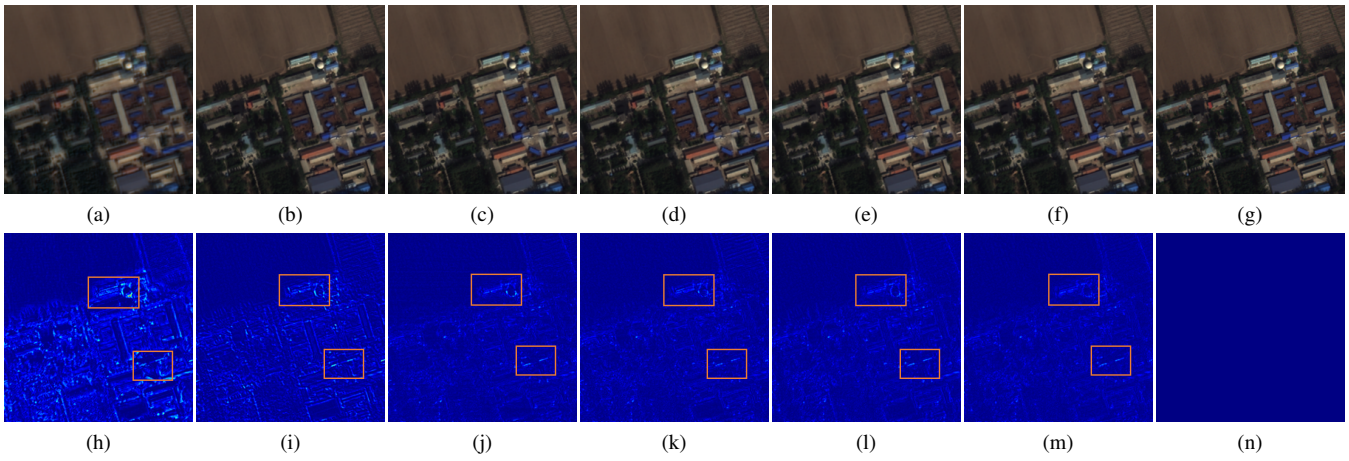

Fig. 13. Pansharpened images resulting from various ablation models. (a) Base. (b) 1SAFEB. (c) 2SAFEBs. (d) 2B_2SA_CI. (e) 2B_2SA_CI_MDC. (f) Proposed. (g) GT. (h)-(n) are the corresponding AEMs of (a)-(g).

TABLE VII
MEAN OBJECTIVE ASSESSMENT ON WORLDVIEW-3 DATASET USING DIFFERENT ABLATION MODELS

Models	SAFEB	CI	MDC	CAB	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q8↑
Base	✗	✗	✗	✗	0.8389	6.7562	5.7776	0.8184	0.8493
1SAFEB	✓	✗	✗	✗	0.8992	5.0044	3.5361	0.8831	0.9021
2SAFEBs	×2	✗	✗	✗	0.9082	4.7607	3.3493	0.8908	0.9107
2B_2SA_CI	×4	✓	✗	✗	0.9204	4.5993	2.7876	0.9265	0.9230
2B_2SA_CI_MDC	×4	✓	✓	✗	0.9333	4.4005	2.6732	0.9401	0.9355
Proposed	×4	✓	✓	✓	**0.9371**	**4.2561**	**2.5979**	**0.9458**	**0.9396**

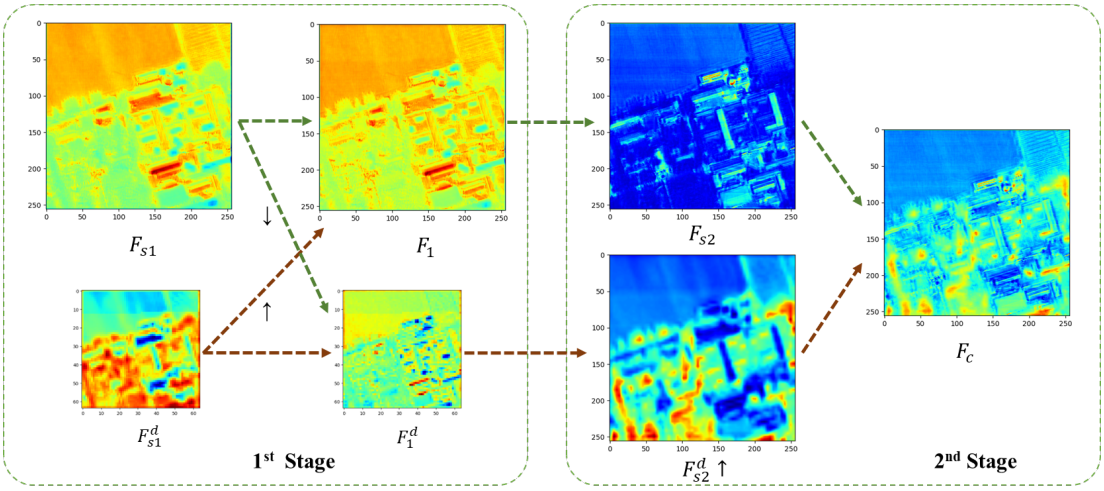

Fig. 14. Visualization of multi-scale feature interaction

TABLE VIII
OBJECTIVE COMPARISON OF THE FUSION RESULTS USING TRANSFORMER AND SEWFORMER

Methods	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q8↑
1B_TF	0.9330	4.4363	2.7126	0.9400	0.9327
1B_SEWF	0.9339	4.3767	2.6971	0.9386	0.9334
2B_TF	0.9365	4.3254	2.6259	0.9449	0.9389
Proposed	**0.9371**	**4.2561**	**2.5979**	**0.9458**	**0.9396**

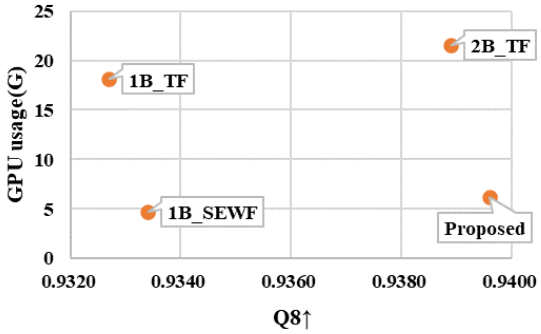

Fig. 15. Comparison of the performance and GPU usage

former, we substitute SEWformer in SAFEB with a Transformer shown in Fig. 1, and evaluate the impact of various encoder configurations. These configurations include a one branch encoder with Transformer (1B_TF), a one branch encoder with SEWformer (1B_SEWF), a two branches interactive encoder with Transformer (2B_TF), and a two branches interactive encoder with SEWformer (Proposed). As subjective fusion results are challenging to differentiate between these models, we primarily rely on objective assessments presented in Table VIII. Notably, the performance of SEWformer-based configurations is slightly superior to their Transformer-based counterparts. Moreover, we recorded GPU usage during the training process using input samples with a batch size of 20 and a cropped image size of 64 × 64. The average Q8 results of the WorldView-3 dataset and the GPU usage of the models using Transformer or SEWformer are illustrated in Fig. 15. It is evident that employing the proposed SEWformer significantly reduces GPU consumption compared to using the Transformer. Thus, the SEWformer is a promising approach that offers relatively high performance while substantially improving GPU utilization.

E. Generalization Experiment

To further validate the universality and generalization ability of our method, we conducted additional cross-dataset experiments. We trained our model, as well as other deep learning-based comparison methods, on the IKONOS dataset, and then evaluated the trained models on the Pléiades dataset. This cross-dataset evaluation helps verify the robustness and adaptability of the proposed approach.

The fusion results are shown in Fig. 16. From the enlarged regions, we can observe that our method demonstrates superior

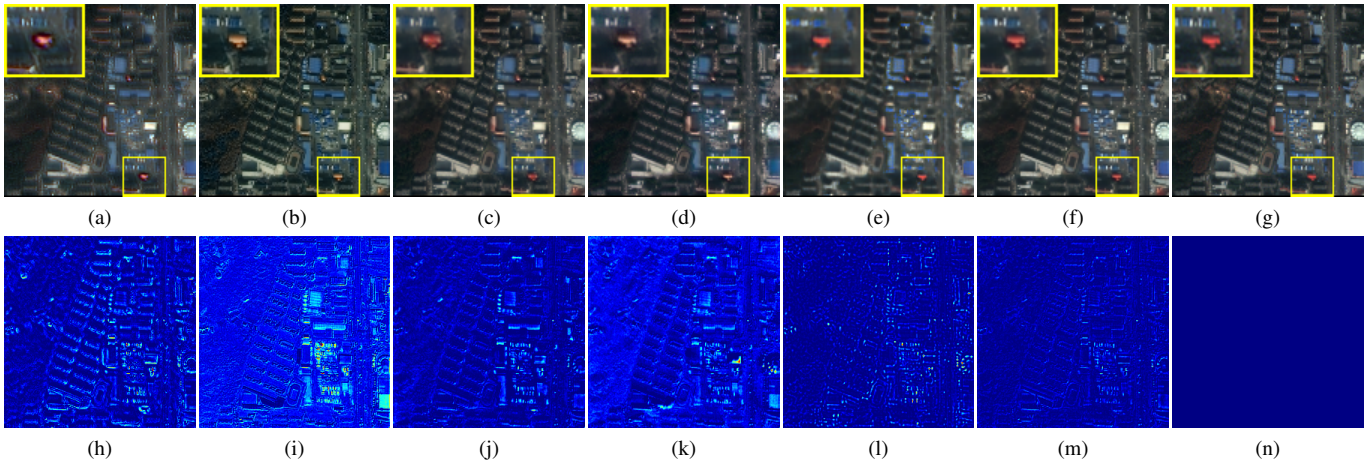

Fig. 16. Fusion results of the images from Pléiades dataset using the model trained on IKONOS dataset. (a) APNN-FT. (b) PCGAN. (c) FusionNet. (d) TDNet. (e) PMRF. (f) Proposed. (g) GT. (h)-(n) are the corresponding AEMs of (a)-(g).

TABLE IX
OBJECTIVE ASSESSMENT OF FUSION RESULTS PRESENTED IN FIG. 16 AND MEAN ASSESSMENT ON PLÉIADES DATASET

Methods	Assessment of Fig. 16					Mean Assessment				
	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑
APPN-FT [30]	0.8160	7.2705	5.4574	0.7471	0.8130	0.8293	5.5122	4.2247	0.7495	0.8221
PCGAN [19]	0.8495	10.1066	6.4227	0.8478	0.8103	0.8553	7.9779	5.5045	0.8343	0.8012
FusionNet [21]	0.9231	3.9069	3.5896	0.8855	0.9197	0.9112	2.7459	2.9133	0.8670	0.9080
TDNet [31]	0.8898	6.6682	4.9121	0.8673	0.8470	0.8893	4.6522	3.8725	0.8626	0.8320
PMRF [25]	0.9002	4.4484	4.3321	0.8236	0.8939	0.8979	3.4240	3.2181	0.8285	0.8853
Proposed	**0.9449**	**3.7265**	**3.0171**	**0.9035**	**0.9462**	**0.9378**	**2.8408**	**2.3179**	**0.8996**	**0.9374**

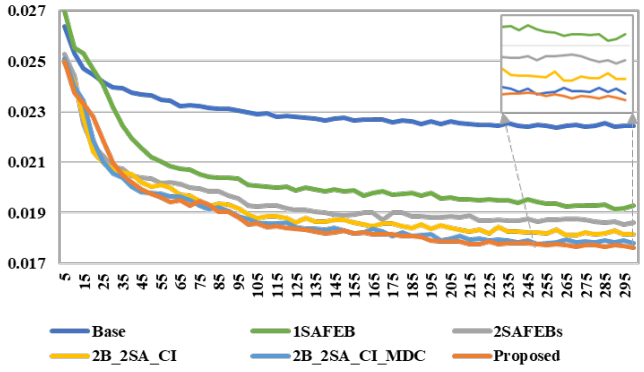

Fig. 17. L1-loss curves of the compared models

TABLE X
EFFICIENCY COMPARISON. NOP REPRESENTS NUMBER OF PARAMETERS.

Methods	Testing Time(s)	FLOPs(G)	NoP(M)
EXP [29]	-	-	-
AWLP-R [9]	0.04(CPU)	-	-
RBDSD [6]	0.03(CPU)	-	-
CDIF [3]	22.76(CPU)	-	-
APPN-FT [30]	3.2(CPU)	-	0.31
PCGAN [19]	0.04(GPU)	11.16	32.62
FusionNet [21]	0.01(GPU)	9.92	0.23
VO+Net [22]	11.65(CPU)	-	0.31
TDNet [31]	0.03(GPU)	19.98	0.49
PMRF [25]	0.04(GPU)	32.62	0.39
Proposed	0.04(GPU)	4.29	0.11

performance in preserving spectral fidelity. The AEMs also indicate that our method produces the fewest residuals.

The objective evaluation presented in Table IX further confirm the advantage of our approach over the comparison methods. Our method achieves significantly better results, showcasing its effectiveness in delivering high-quality pan-sharpened images across different datasets.

F. Convergence Analysis

To gain insights into the training dynamics and stability of the proposed models, we conducted a convergence analysis using the L1 loss as the metric. During the training process, we recorded the L1 loss every five epochs to track the convergence. The resulting loss curves are presented in Fig. 17. Upon analyzing the curves, we observe that the ablation models exhibit more fluctuating curves or higher L1 loss values. In contrast, our proposed method demonstrates a stable convergence pattern with a minimum L1 loss. This indicates that our method efficiently learns and emphasizes the important spatial and spectral attentions, leading to improved performance.

G. Efficiency Analysis

We conducted an efficiency analysis to evaluate the computational efficiency of our method. We measured the testing time, floating-point operations (FLOPs), and the number of parameters (NoP) required by each comparison method to

process a single input sample consisting of the 1 × 1 × 256 × 256 and 1 × 4 × 64 × 64 tensors. These measurements were performed on the same hardware platform, including an RTX3090 GPU and 24GB of RAM, ensuring a fair comparison between the models. It is worth noting that we only calculate model size and FLOPs when using GPU. The comparison results are presented in Table X. As observed from the table, our method achieved the lowest FLOPs, and fewest NoP compared to all the other DL-based methods included in the comparison. The testing time is similar across different DL-based methods when utilizing GPU acceleration. These results indicate the superior computational efficiency of our approach. This efficiency can be primarily attributed to the design of the SEWformer in the encoder and the utilization of multiscale dilation convolution in the decoder, which effectively conserve computational resources. Combining the low GPU usage as displayed in Fig. 15, our method stands as a lightweight approach with high performance.

The low computational requirements of our model, as evidenced by the low levels of FLOPs and NoP, make it suitable for deployment on a variety of hardware platforms. From resource-constrained edge devices to powerful server-grade GPUs, the model can be effectively utilized across different computing setups.

The model's ability to deliver near real-time pansharpening performance on mobile or embedded platforms could enable its use in time-sensitive applications, such as disaster response or urban planning. Alternatively, the model's efficiency on high-performance servers makes it suitable for large-scale satellite image processing and analysis tasks in remote sensing or geospatial intelligence domains.

This versatility in hardware adaptability, combined with the model's robust pansharpening capabilities, highlights its potential for practical deployment across a wide range of real-world applications.

H. Classification Application

In this section, we focus on the classification application of our proposed method and evaluate its efficacy in solving classification tasks. The classification task was performed using the ENVI classification tool, and the K-means classification algorithm is employed to classify objects in the images. The classification results were compared with GT to assess whether pansharpening improved the accuracy of object classification.

Taking the fusion results depicted in Fig. 5 as an example, the corresponding classification results are presented in Fig. 18. The figure illustrates the superior classification accuracy achieved by our proposed method when compared to GT, particularly in the red and blue regions denoting concavity and protrusion respectively.

To provide a more comprehensive evaluation of the classification results, we utilize several well-established evaluation metrics, including kappa coefficient (K↑), commission error (CE↓), overall accuracy (OA↑), and omission error (OE↓), as shown in Table XI. It can be observed that our method consistently outperforms the other methods in all metrics, exhibiting a significant advantage in terms of classification accuracy. We can conclude that our method holds great potential for enhancing object classification accuracy in remote sensing applications.

I. More discussion

One of the key advantages of our method is its superior pansharpening performance, as demonstrated in both subjective and objective evaluations in the experimental results. The two-stage SAFEB and the integration of SEWformer, cross-scale interaction, and multiscale dilation convolution have enabled our model to effectively capture and fuse multi-scale features, leading to highly detailed and spectrally-accurate pansharpened results. Additionally, our method has exhibited

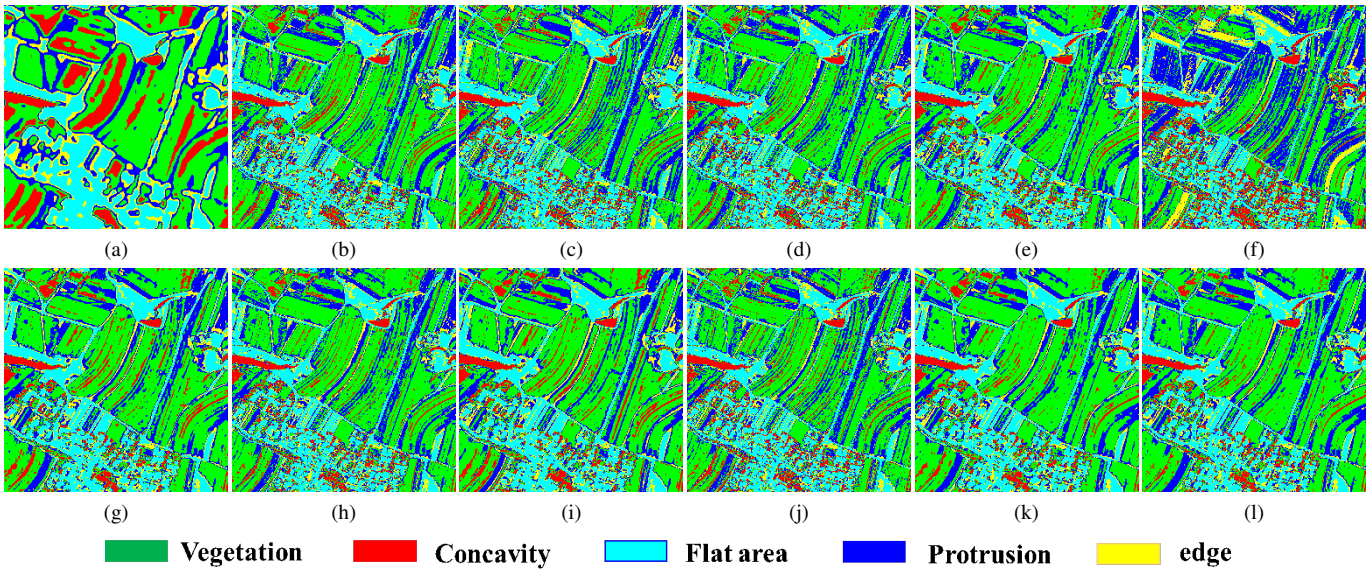

Fig. 18. Classification results for the images shown in Fig. 5. (a) EXP. (b) AWLP-R. (c) RBDSD. (d) CDIF. (e) APNN-FT. (f) PCGAN. (g) FusionNet. (h) VO + Net. (i) TDNet. (j) PMRF. (k) Proposed. (l) GT.

TABLE XI
OBJECTIVE ASSESSMENT OF THE CLASSIFICATION RESULTS IN FIG.18

Methods	OA↑	K↑	CE↓	OE↓
EXP [29]	0.7022	0.6204	0.2920	0.2903
AWLP-R [9]	0.7634	0.6991	0.3037	0.2744
RBDSD [6]	0.7435	0.6414	0.3188	0.2797
CDIF [3]	0.7890	0.7023	0.2722	0.2471
APNN-FT [30]	0.7939	0.7080	0.2659	0.2479
PCGAN [19]	0.5277	0.3674	0.4680	0.4454
FusionNet [21]	0.8275	0.7565	0.2304	0.2013
VO+Net [22]	0.8081	0.7293	0.2488	0.2240
TDNet [31]	0.7922	0.7085	0.2745	0.2311
PMRF [25]	0.7559	0.6590	0.2803	0.2686
Proposed	**0.8773**	**0.8251**	**0.1593**	**0.1546**

strong generalization capabilities, maintaining robust performance when evaluated on diverse datasets. This cross-dataset validation underscores the universality and adaptability of our approach. Furthermore, the computational efficiency of our model, reflected in its low FLOPs and parameter count, allows for its deployment on a diverse range of hardware platforms. As a result, our method is well-suited for real-world applications, accommodating varying sensor and hardware conditions.

While the proposed method has shown promising results, there are a few limitations. For example, our method relies on a large amount of training data to achieve optimal performance. The method may not be well-suited for scenarios with limited sample availability. Besides, the current computational efficiency of the model, though relatively high, may have room for further optimization.

V. CONCLUSION

To address the challenges of scale inconsistency, high computational cost, and insufficient attention representation in pansharpening, this paper introduces a lightweight network that leverages cross-scale interaction and spatial-spectral enhanced window attention. The proposed SEWformer integrates extra spatial-spectral linear attention to enhance attention information within and between interleaved windows. A cross-scale interactive encoder is developed based on the two-stage SEWformers to improve scale consistency in the fusion process. Furthermore, a decoder combining a multiscale residual network with channel attention accurately extracts image details. Extensive experimental validation on multiple datasets and an ablation study demonstrate the superior performance and efficiency of our method compared to state-of-the-art techniques. Additionally, a classification experiment conducted as part of our evaluation highlights the significant potential of our approach in classification applications. Future research will focus on further enhancing the computational efficiency of our method, exploring few-shot learning or meta-learning techniques to improve its adaptability to small-sample settings, and investigating the application of our approach in other remote sensing tasks beyond pansharpening.

REFERENCES

[1] H. Lu, Y. Yang, S. Huang, X. Chen, B. Chi, A. Liu, and W. Tu, "AWFLN: An Adaptive Weighted Feature Learning Network for Pansharpening," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–15, 2023.

[2] L.-j. Deng, G. Vivone, M. E. Paoletti, G. Scarpa, J. He, Y. Zhang, J. Chanussot, and A. Plaza, "Machine Learning in Pansharpening: A benchmark, from shallow to deep networks," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 279–315, Sep. 2022.

[3] 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, pp. 1–15, 2022.

[4] Q. Cao, L.-J. Deng, W. Wang, J. Hou, and G. Vivone, "Zero-shot semi-supervised learning for pansharpening," *Information Fusion*, vol. 101, p. 102001, Jan. 2024.

[5] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS +pan data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.

[6] 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.

[7] X. Zhou, J. Liu, S. Liu, L. Cao, Q. Zhou, and H. Huang, "A GIHS-based spectral preservation fusion method for remote sensing images using edge restored spectral modulation," *ISPRS J. Photogram. Remote Sens.*, vol. 88, pp. 16–27, Feb. 2014.

[8] Y. Yang, H. Lu, S. Huang, and W. Tu, "Remote Sensing Image Fusion Based on Fuzzy Logic and Salience Measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1943–1947, Dec. 2019.

[9] G. Vivone, L. Alparone, A. Garzelli, and S. Lolli, "Fast Reproducible Pansharpening Based on Instrument and Acquisition Modeling: AWLP Revisited," *Remote Sensing*, vol. 11, no. 19, p. 2315, 2019.

[10] G. Vivone, S. Marano, and J. Chanussot, "Pansharpening: Context-Based Generalized Laplacian Pyramids by Robust Regression," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6152–6167, 2020.

[11] Y. Yang, C. Wan, S. Huang, H. Lu, and W. Wan, "Pansharpening Based on Low-Rank Fuzzy Fusion and Detail Supplement," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 13, pp. 5466–5479, Sep. 2020.

[12] H. Lu, Y. Yang, S. Huang, and W. Tu, "An Efficient Pansharpening Approach Based on Texture Correction and Detail Refinement," *IEEE Geosci. Remote Sensing Lett.*, vol. 19, pp. 1–5, 2022.

[13] Y. Yang, H. Lu, S. Huang, and W. Tu, "Pansharpening Based on Joint-Guided Detail Extraction," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 14, pp. 389–401, 2021.

[14] T. Wang, F. Fang, F. Li, and G. Zhang, "High-Quality Bayesian Pansharpening," *IEEE Trans. on Image Process.*, vol. 28, no. 1, pp. 227–239, Jan. 2019.

[15] 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," *Information Fusion*, vol. 52, pp. 76–89, Dec. 2019.

[16] Y. Yang, H. Lu, S. Huang, W. Wan, and L. Li, "Pansharpening Based on Variational Fractional-Order Geometry Model and Optimized Injection Gains," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 15, pp. 2128–2141, 2022.

[17] 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. on Image Process.*, vol. 31, pp. 918–933, 2022.

[18] J. Yang, X. Fu, Y. Hu, H. Yue, and J. Paisley, "PanNet: A Deep Network Architecture for Pan-Sharpening," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[19] F. Ozcelik, U. Alganci, E. Sertel, and G. Unal, "Rethinking CNN-Based Pansharpening: Guided Colorization of Panchromatic Images via GANs," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 4, pp. 3486–3501, Apr. 2021.

[20] W. Tu, Y. Yang, S. Huang, W. Wan, L. Gan, and H. Lu, "MMDN: Multi-Scale and Multi-Distillation Dilated Network for Pansharpening," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–14, 2022.

[21] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail Injection-Based Deep Convolutional Neural Networks for Pansharpening," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 8, pp. 6995–7010, 2021.

[22] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J.-F. Hu, and G. Vivone, "VO+Net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[23] Y. Shi, A. Tan, N. Liu, W. Li, R. Tao, and J. Chanussot, "A Pansharpening Method Based on Hybrid-Scale Estimation of Injection Gains," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–15, 2023.

[24] X. Meng, N. Wang, F. Shao, and S. Li, "Vision Transformer for Pansharpening," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–11, 2022.
[25] L. Hou, B. Zhang, and B. Wang, "PAN-Guided Multiresolution Fusion Network Using Swin Transformer for Pansharpening," *IEEE Geosci. Remote Sensing Lett.*, vol. 20, pp. 1–5, 2023.
[26] X. Su, J. Li, and Z. Hua, "Transformer-Based Regression Network for Pansharpening Remote Sensing Images," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–23, 2022.
[27] W. G. C. Bandara and V. M. Patel, "HyperTransformer: A Textural and Spectral Feature Fusion Transformer for Pansharpening," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 1757–1767.
[28] F. Dadrass Javan, F. Samadzadegan, S. Mehravar, A. Toosi, R. Khatami, and A. Stein, "A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery," *ISPRS J. Photogram. Remote Sens.*, vol. 171, pp. 101–117, Jan. 2021.
[29] 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, Jan. 2002.
[30] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-Adaptive CNN-Based Pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.
[31] T.-J. Zhang, L.-J. Deng, T.-Z. Huang, J. Chanussot, and G. Vivone, "A Triple-Double Convolutional Neural Network for Panchromatic Sharpening," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 34, no. 11, pp. 9088–9101, Nov. 2023.
[32] H. Lu, Y. Yang, S. Huang, X. Chen, H. Su, and W. Tu, "Intensity mixture and band-adaptive detail fusion for pansharpening," *Pattern Recognition*, vol. 139, p. 109434, 2023.
[33] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, and J. Chanussot, "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.