Invertible Attention-Guided Adaptive Convolution and Dual-Domain Transformer for Pansharpening

Qun Song, Hangyuan Lu, Chang Xu, Rixian Liu, Weiguo Wan, Wei Tu

***Abstract*—Pansharpening is the process of fusing a multispectral (MS) image with a panchromatic (PAN) image to produce a high-resolution multispectral (HRMS) image. However, existing techniques face challenges in integrating long-range dependencies to correct locally misaligned features, which results in spatial-spectral distortions. Moreover, these methods tend to be computationally expensive. To address these challenges, we propose a novel detail injection algorithm and develop the invertible attention-guided adaptive convolution and dual-domain Transformer (IACDT) network. In IACDT, we designed an invertible attention mechanism embedded with spectral-spatial attention to efficiently and losslessly extract locally spatial-spectral-aware detail information. Additionally, we presented a frequency-spatial dual-domain attention mechanism that combines a frequency-enhanced Transformer and a spatial window Transformer for long-range contextual detail feature correction. This architecture effectively integrates local detail features with long-range dependencies, enabling the model to correct both local misalignments and global inconsistencies. The final HRMS image is obtained through a reconstruction block that consists of residual multi-receptive field attention. Extensive experiments demonstrate that IACDT achieves superior fusion performance, computational efficiency, and outstanding results in downstream tasks compared to state-of-the-art methods. The code is available at https://github.com/yotick/IACDT-pansharpening.**

***Index Terms*—Pansharpening, dual-domain, Transformer, adaptive convolution**

I. Introduction

Remote sensing image fusion plays a crucial role in various applications, such as land cover classification, change detection, and environmental monitoring [1], [2]. It involves integrating complementary information from multiple remote sensing images including panchromatic (PAN) and multispectral (MS) images. PAN images capture the scene with a single broad-spectrum band, providing high spatial resolution but limited spectral information. On the other hand, MS images consist of multiple spectrum bands, offering rich spectral information but at a lower spatial resolution. By fusing these two types of data, the resultant high-resolution MS (HRMS) image enhances both the spatial details and spectral fidelity, surpassing the limitations of individual input images. The HRMS image enhances the interpretability, spatial details, and spectral fidelity compared to individual input images, enabling improved analysis and decision-making [3], [4]. The fusion process is also called pansharpening.

Over the years, numerous image fusion techniques have been proposed to address the challenges associated with pansharpening. These techniques can be broadly categorized into traditional methods and data-driven methods [5]. Traditional methods include a variety of approaches, including component substitution (CS), multi-resolution analysis (MRA), and Model-based methods [6]. The CS method focuses on substituting the intensity component of the MS image with the high-resolution PAN image while preserving the spectral information. Popular CS methods include Gram-Schmidt adaptive(GSA) approach [7], robust band-dependent spatial-detail (RBDSD) [8], generalized intensity–hue–saturation (GIHS) transform [9], etc. The MRA method, on the other hand, relies on a multi-resolution decomposition of both the PAN and MS images using techniques like wavelet or pyramid transforms. The revised additive wavelet luminance proportional (AWLP-R) [10], generalized Laplacian pyramid [11], and adaptive multiscale bilateral filtering [12] are the recently advanced MRA-based methods. However, the CS- and MRA- based methods can hardly balance the spatial and spectral qualities [13].

As an alternative traditional approach, researchers tend to develop mathematical models that capture the statistical properties, spectral correlations, and spatial dependencies of PAN and MS images. These methods utilize regression models [14], Bayesian frameworks [15], or Markov random fields [16] to estimate the fused image by optimizing certain criteria or constraints. In addition, Wen et al [17] introduced LNM-PS, a pansharpening method that incorporates a learnable nonlinear mapping into the spatial fidelity term and achieve impressive performance across various datasets. Model-based fusion techniques leverage the inherent characteristics of the data, offering the potential for enhanced fusion results. However, the quality of fusion heavily relies on the accuracy of the model and the appropriate setting of parameters [18].

Deep learning-based pansharpening methods are popular due to their ability to learn complex and non-linear relationships between the input images [19]. Convolutional Neural Networks (CNNs) have been widely employed in these methods to capture spatial and spectral features for effective

This work is supported by the National Natural Science Foundation of China (No.62362035, No.62261025, and No.62361030) *(Hangyuan Lu and Qun Song contributed equally to this work.)(Corresponding author: Hangyuan Lu, Chang Xu.)*

S. Song, H. Lu, and R. Liu are with the College of Information Engineering, and also with Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua University of Vocational Technology, Jinhua 321007, China (e-mail: 20050178@jhc.edu.cn, lhyhziee@163.com; lynxu223@163.com).

Chang xu is with the Hangzhou Consumer Council, Hangzhou 310008, China (e-mail: yotic27@gmail.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)

fusion. For example, Ozcelik et al. [20] treated pansharpening as an image colorization task and proposed a self-supervised GAN framework called PCGAN, demonstrating promising spatial quality. Zhang et al. [21] proposed a dual-task collaborative promotion network (DCPNet) for pansharpening, which integrates LRMS super-resolution reconstruction and pansharpening tasks to achieve joint optimization in spectral-spatial qualities. To enhance interpretability in deep learning networks, an alternative strategy involves the integration of CNN with traditional techniques [22]. For instance, Deng et al. [23] introduced detail injection-based deep convolutional neural networks called FusionNet, which combines CNN with traditional fusion schemes to estimate non-linear injection details. Wu et al. [24] presented VO+Net, which enhances the deep learning framework by integrating spatial and spectral fidelity terms, supplemented by a weighted regularization term. Further, Wang et al [25] proposed VOGTNet, a variational optimization-guided two-stage network for robust multispectral pansharpening, effectively addressing noise and blur while improving image quality.

Recently, Transformers have gained significant attention for their ability to model long-range dependencies and capture global context effectively. Su et al. [26] proposed a transformer-based regression network for pansharpening, which effectively extracts global spectral information and spatial details. Zhang et al. [27] combined convolutional neural networks and transformers to explore common information and reduce redundancy in deep feature extraction. To improve efficiency, the Swin Transformer was developed, utilizing a hierarchical structure that enables efficient and scalable processing of large images. For example, Hou et al. [28] proposed a PAN-guided multiresolution fusion (PMRF) network based on Swin Transformer to enhance spatial resolution and feature representation, demonstrating superior performance in terms of detail preservation.

While significant advancements have been achieved in pansharpening techniques, several challenges still need to be tackled. Firstly, most remote sensing images from different modal sensors are not strictly aligned, especially in the certain local areas, as illustrated in Fig. 1. By using the horizontal alignment lines, we can observe local misalignment between the up sampled MS (UPMS) and PAN images. Furthermore, structural similarity analysis of the source images reveals significant local structural differences between them. However, existing methods primarily focus on local feature extraction, struggling to effectively integrate long-range contextual information, which leads to spatial-spectral distortions in the fused image. Secondly, improving the interpretability of neural networks remains a challenge. Lastly, current approaches are computationally demanding and often rely on increased complexity to enhance performance, limiting their practical applicability.

To address these challenges, we propose a novel detail injection algorithm that integrates local-global joint detail optimization, and construct an efficient network called invertible attention-guided adaptive convolution and dual-domain Transformer (IACDT) based on this algorithm. The network introduces invertible attention mechanism that information-lossless integrating pixel attention, channel attention and spatial attention, thereby effectively guide the adaptive kernels to extract local detailed information. Furthermore, we develop a frequency-enhanced attention (FEA) mechanism to capture salient frequency components. Combined with spatial window attention (SWA), this dual-domain attention facilitates effective global feature correction. The extracted details are further refined by incorporating the adaptive convolution module and dual-domain attention within a multiscale residual architecture. These mechanisms allows the model to correct locally misaligned features while maintaining long-range contextual consistency. Further, we carefully design tailored loss function, including content, spatial, and perceptual losses, to guide the generation of fused images. The fusion performance comparison on the WorldView-3 dataset with state-of-the-art (SOTA) deep learning-based methods, such as FusionNet [23], TDNet [29], and PMRF [28], is presented in Fig. 2. We employ the widely adopted Q8↑ metric to evaluate the overall quality, and report the floating-point operations (FLOPs) and the number of parameters (NoP) to assess computational efficiency. As evident from the results, our proposed method achieves superior fusion quality while maintaining computational efficiency, outperforming the other SOTA techniques in both aspects. To sum up, the main contributions of our approach is as follows:

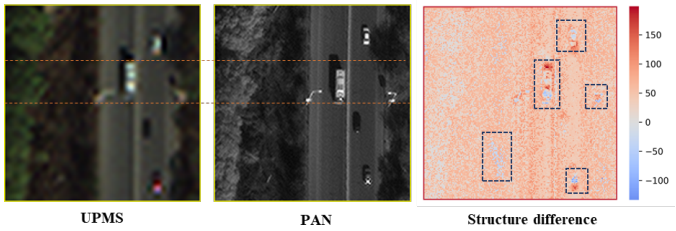


Fig. 1. Alignment comparison between UPMS and PAN images.

1. A novel detail-injection algorithm is proposed based on local-global joint detail optimization, and an efficient pansharpening network, termed IACDT, is developed accordingly. This network enhances fusion performance with high efficiency while improving interpretability.

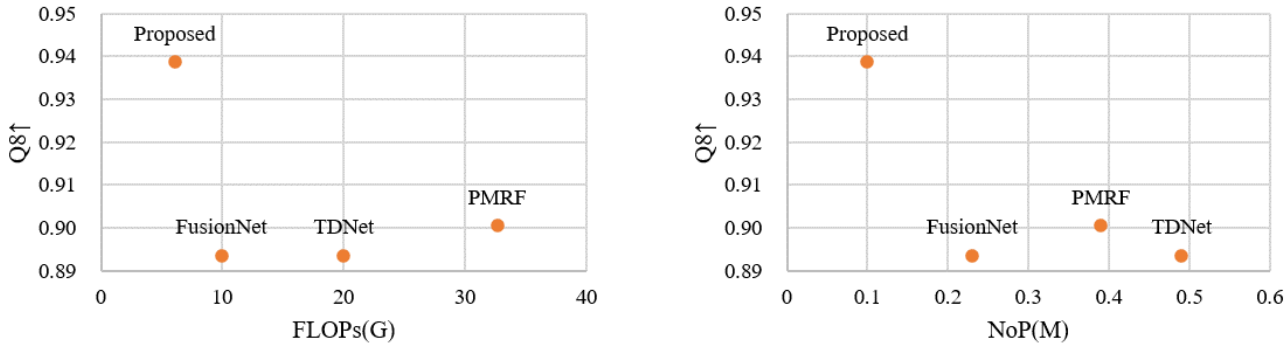
2. An invertible attention-guided adaptive convolution (IAAC) module is designed to adaptively adjust convolutional kernels based on spatial-spectral characteristics. This module efficiently enhances the extraction of local details while losslessly and accurately preserving critical spatial and spectral features.

3. A frequency-spatial dual-domain attention (FSDA) module is presented, consisting of FEA and SWA with in Transformer architectures. This module effectively integrates long-range contextual information and further correct the extracted details.

II. Related Work

A. Wavelet Transform

The wavelet transform (WT) is a powerful mathematical tool for multi-resolution analysis of signals and images. It decomposes a signal into a set of basis functions, known as



(a) (b)

Fig. 2. Performance comparison with the SOTA deep-learning based methods

wavelets, which are localized in both time space and frequency domains [30]. The discrete wavelet transform (DWT) is particularly suitable for digital image processing, where an image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times C}$ is decomposed into four subbands at each level:

$$\boldsymbol{I}_{LL}, \boldsymbol{I}_{LH}, \boldsymbol{I}_{HL}, \boldsymbol{I}_{HH} = \mathrm{DWT}(\boldsymbol{I}), \tag{1}$$

where $\boldsymbol{I}_{LL}$ represents the low-frequency approximation coefficients, while $\boldsymbol{I}_{LH}$, $\boldsymbol{I}_{HL}$, and $\boldsymbol{I}_{HH}$ correspond to the high-frequency detail coefficients in the horizontal, vertical, and diagonal directions, respectively. Each subband is downsampled by a factor of 2 along each dimension, effectively capturing features at different scales and orientations.

The multiresolution and localization properties of WT make it well-suited for pansharpening, as it can effectively integrate the spatial details from the high-resolution PAN image while preserving the spectral information from the MS image. Some researchers operate pansharpening network in the frequency domain by integrating wavelet transform to decompose the low-resolution input into different frequency bands, and use the CNN network to predict the high-frequency components [31]. This approach has shown promising results in remote sensing applications.

B. Detail Injection Models

Detail injection models represent a widely adopted class of pansharpening techniques that aim to inject the high-frequency details from the PAN image into the MS image. These models typically involve decomposing the input images into approximation and detail components, followed by injecting the extracted details into the MS image.

One popular approach is the CS-based method, which injects high-frequency details from the PAN image into the multispectral image using component replacement:

$$\boldsymbol{I}_{fused} = \boldsymbol{I}_{ms\uparrow} + g\left(\boldsymbol{I}_{pan} - \boldsymbol{S}_{ms\uparrow}\right), \tag{2}$$

where $\boldsymbol{I}_{fused}$, $\boldsymbol{I}_{ms\uparrow}$, and $\boldsymbol{I}_{pan}$ represent fused, UPMS, and PAN images. $\boldsymbol{S}_{ms\uparrow}$ is the spatial component of the UPMS image, g is an injection coefficient.

Another approach employs MRA tools like wavelets or contourlets to extract and inject details at multiple scales and orientations [5]. For instance, the AWLP method decomposes the PAN and MS images using an à trous wavelet transform and injects the details based on a luminance-proportional model [10]:

$$\boldsymbol{I}^{L}_{fused} = \boldsymbol{I}^{L}_{ms\uparrow} + g\left(\boldsymbol{I}^{L}_{pan} - \boldsymbol{I}^{L}_{ms\uparrow}\right), \tag{3}$$

$$\boldsymbol{I}^{H}_{fused} = \boldsymbol{I}^{H}_{ms\uparrow} + g\left(\boldsymbol{I}^{H}_{pan}\right), \tag{4}$$

where $\boldsymbol{I}^{L}$ and $\boldsymbol{I}^{H}$ denote the low- and high-frequency wavelet coefficients, respectively.

Both of MRA- and CS- based approaches can be regarded as the detail injection model [6]. The general form of the model can be expressed as:

$$\boldsymbol{I}_{fused} = \boldsymbol{I}_{ms\uparrow} + g \cdot D_e \quad s.t.\ D_e = f_d(\boldsymbol{I}_{ms\uparrow}, \boldsymbol{I}_{pan}), \tag{5}$$

where D_e represents extracted details, and $f_d(\cdot)$ represents the detail extraction function. The detail injection model have been widely studied and employed in various pansharpening applications due to their simplicity, effectiveness, and ability to preserve spectral information while enhancing spatial details [32]. Building on this approach, Wang et al [33] proposed LRTCP, a pansharpening method that integrates haze correction with low-rank tensor completion to enhance HRMS image reconstruction. Furtehrmore, Wu et al [34] proposed implicit neural feature fusion function in detail injection scheme, leveraging dual high-frequency fusion and a parameter-free cosine similarity method to achieve impressive performance on pansharpening tasks.

III. Proposed Method

A. Overall Framework

As stated in Equation (5), achieving high-quality fusion results depends on effective detail extraction and the proper setting of the injection coefficient. Denoting D_{ref} as the reference detail, and D_e^{Ω} and D_{ref}^{Ω} as the local details in region Ω corresponding to D_e and D_{ref}, respectively, To estimate accurate details, it is essential to perform detail extraction in the local domain, thereby effectively capturing and adapting to the diverse local texture characteristics of the input images. As the extracted local details D_{Ω}^{e} and the reference local details

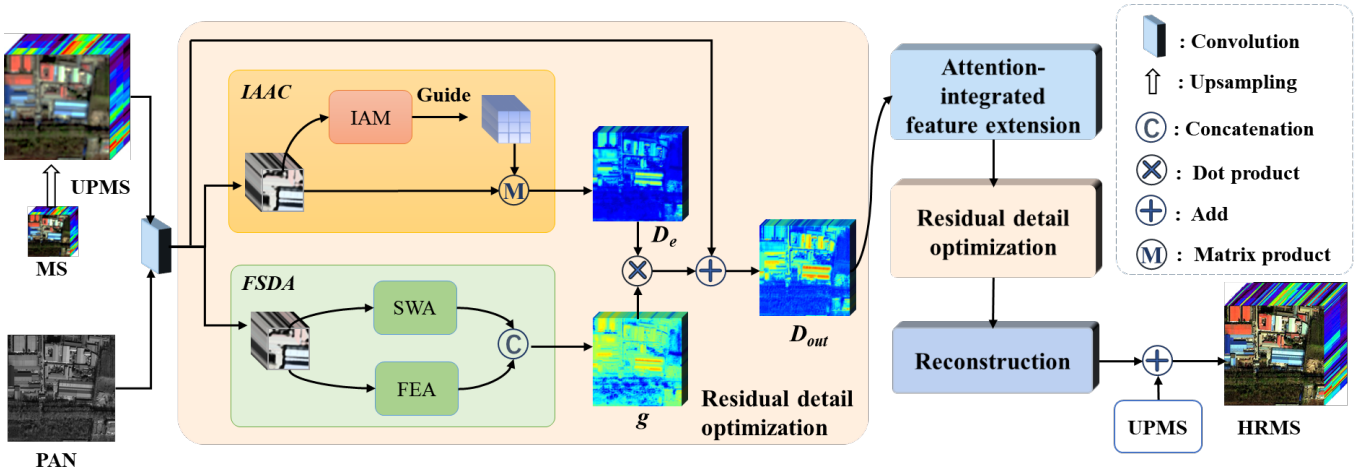


Fig. 3. Architecture of the IACDT network. IAAC: spatial-spectral attention-guided adaptive convolution. FSDA: frequency-spatial dual-domain attention. SWA: spatial window attention. FEA: frequency enhanced attention.

D_{Ω}^{ref} typically exhibit a linear relationship [18], [35], this relationship can be expressed as:

$$D_{ref}^{\Omega} = \alpha_{\Omega} D_e^{\Omega} + \beta_{\Omega}, \tag{6}$$

where α_{Ω} is the local mapping coefficient on the specific local region Ω, and β_{Ω} is the local error. To optimize the balance between spatial and spectral qualities, we define the coefficient α_{Ω} based on local spatial attention A_{spat}^{Ω} and spectral attention A_{spec}^{Ω}. Thus, by omitting the error parameter and assuring that the extracted details are close to the reference details, i.e., $D_e \approx D_{ref}$, the extracted detail is defined as follows.

$$D_e = \sum\nolimits_{\Omega} \alpha_{\Omega} D_{\Omega}^{e} \quad s.t.\ \alpha_{\Omega} = f_{\alpha}(A_{spat}^{\Omega}, A_{spec}^{\Omega}), \tag{7}$$

where $f_{\alpha}(\cdot)$ denotes a function used to extract local attention.

The injection coefficient g also holds significant importance in determining the fusion quality. To enhance the locally extracted details by integrating global contextual information, this paper constructs the coefficient g through a frequency-spatial dual domain global attention. Specifically, we formulate g as a function of $g = f_g(A_{freq}^{g}, A_{spat}^{g})$, where A_{freq}^{g} and A_{spat}^{g} represent the frequency-domain and spatial-domain attention maps, respectively. Combining Equations (5) and (7), the process of extracting and optimizing details is defined as:

$$\begin{aligned} &D_{out} = g \cdot D_e = g\left(\textstyle\sum_{\Omega} \alpha_{\Omega} D_{\Omega}^{e}\right) \\ &s.t.\ g = f_g(A_{freq}^{g}, A_{spat}^{g}), \\ &\alpha_{\Omega} = f_{\alpha}(A_{spat}^{\Omega}, A_{spec}^{\Omega}), \\ &D_{\Omega}^{e} = f_d(\mathbf{I}_{ms\uparrow}, \mathbf{I}_{pan}). \end{aligned} \tag{8}$$

Based on Equation (8), we design the IACDT network as shown in Fig. 3. In the network, the initial features are obtained by concatenating the source image and applying a convolution operation. These features are then fed into the residual detail optimization block. This block is mainly composed of two components: IAAC module which estimates D_e, and FSDA module which computes g. To enable multiscale feature processing, the attention-integrated feature extension block is designed to expand the channel dimension, allowing the subsequent residual detail optimization block to operate on features at multiple scales. The output from the extended detail optimization block is subsequently passed into a reconstruction block, which incorporates residual multi-receptive fields attention. The final HRMS image is then obtained by injecting the reconstructed details into the UPMS image.

B. IAAC Module

The IAAC module utilizes both spectral and spatial information to guide the adaptive convolution, facilitating more efficient extraction of local details. The flowchart of IAAC module is shown in Fig. 4. Specifically, the local initial feature, denoted as F_{in}^{Ω}, is first processed by a convolution with output channel of k^2. To fully capture spectral-spatial aware detail information, the output features then undergo a carefully designed invertible attention mechanism (IAM). The output of the IAM is flattened to $k \times k$ size and repeated along the channel dimension to guide and optimize the convolutional weights. Finally, the optimized adaptive weights are matrix-multiplied with the unfolded input feature F_{in} to obtain the detail information D_e. Mathematically, it can be expressed as:

$$\begin{aligned} &\alpha_{\Omega} = \mathrm{Re}(Conv(Cat(A_{spat}^{\Omega}, A_{spec}^{\Omega}))) \otimes W_A, \\ &D_e = \alpha_{\Omega} * UF(F_{in}), \end{aligned} \tag{9}$$

where $\otimes$ and * denote element-wise multiplication and matrix multiplication, respectively. $Cat(\cdot)$, $Conv(\cdot)$, $\mathrm{Re}(\cdot)$, and $UF(\cdot)$ represent concatenation function, convolution layer, repeat operation, and unfold operation, respectively. W_A denotes adaptive convolution weight. By jointly leveraging spectral and spatial information to guide the convolution process, the IAAC module can efficiently extract required local details.

Invertible neural operators offer advantages such as efficient training and stable optimization. Building on this framework, we propose IAM, which incorporates pixel attention, spatial attention, and spectral attention mechanisms into the invertible architecture. These mechanisms effectively enhance the model's capability to preserve and refine spectral-spatial features while maintaining computational efficiency. As is shown in Fig. 5. The input feature is initially processed

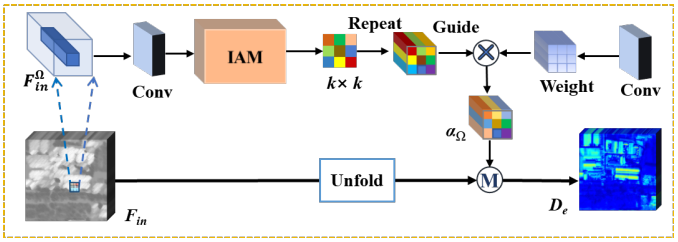


Fig. 4. Flowchart of IAAC module. IAM represents invertible attention mechanism.

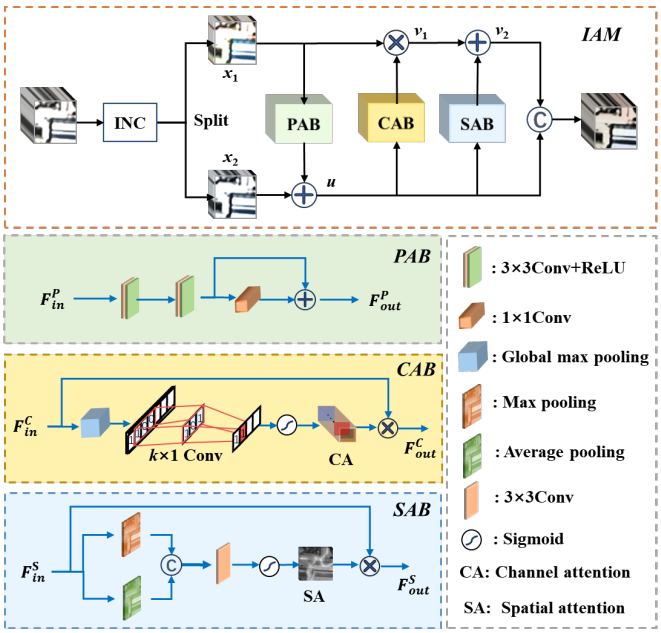


Fig. 5. Architecture of IAM. INC: invertible convolution. PAB: pixel attention block. CAB: channel attention block. SAB: spatial attention block.

through an invertible convolution (INC) layer, a lightweight yet expressive operation that guarantees invertibility. This ensures seamless information flow in both forward and reverse directions without any loss. Subsequently, the feature is split into two components, x_1 and x_2, to facilitate efficient information propagation. To capture salient pixel information while preserving spatial-spectral fidelity, we construct pixel attention block (PAB), channel attention block (CAB), and spatial attention block (SAB) within the IAM. Denoting the process of PAB, CAB, and SAB as $f_P(\cdot)$, $f_C(\cdot)$, and $f_S(\cdot)$, respectively, then the forward flow of IAM can be expressed as:

$$\begin{cases} u = f_P(x_1) + x_2, \\ v_1 = \theta \cdot \left(\sigma\left(f_C(u)\right) \times 2 - 1\right), \\ v_2 = x_1 \cdot \exp(v_1) + f_S(u), \\ O_{IAM} = Cat(u, v_2), \end{cases} \tag{10}$$

where θ represents the scale coefficient, O_{IAM} represents the output of IAM. $\sigma(\cdot)$ denotes the sigmoid activation function. The backward flow corresponds to the reverse process of Eq. (10).

The detailed structures of PAB, CAB, and SAB are shown in Fig. 5. Specifically, the PAB consists primarily of 3×3 convolution layers, ReLU activations, and a residual 1×1 convolution. CAB consists of an adaptive max pooling (AMP) layer to squeeze the spatial dimensions, followed by a one-dimensional convolution (OC) and sigmoid activation to efficiently produce the spectral attention weights. The CAB plays the role of A_{spec}^{Ω} in Eq. (8), and the process is expressed as:

$$F_{out}^{C} = \sigma\left(OC(GMP(F_{in}^{C}))\right) \otimes F_{in}^{C}, \tag{11}$$

On the other hand, the SAB involves employing a max pooling (MP) and average pooling (AP) convolutional layer, followed by sigmoid activation, to generate spatial attention weights. The SAB plays the role of A_{spat}^{Ω} in Eq. (8) and can be expressed as:

$$F_{out}^{S} = \sigma(Conv(Cat(MP(F_{in}^{S}), AP(F_{in}^{S})))) \otimes F_{in}^{S}. \tag{12}$$

C. FSDA Module

The FSDA module serves as the injection coefficient g in Eq. (8). To globally correct the extracted details, the FSDA module is designed by integrating two parallel attention mechanisms: frequency-enhanced attention and spatial window attention, corresponding to A_{freq}^{g} and A_{spat}^{g} in Eq. (8), respectively, as illustrated in Fig. 6.

For frequency enhanced attention, to fully exploit the advantage of DWT in extracting global frequency information, the input feature is first decomposed into a low-frequency component (LL) and three high-frequency components (LH, HL, HH) using DWT. To reinforce the salient frequency components, these four components are then enhanced through a depth-wise separable convolution (DSC) block which includes DSC, batch norm (BN), and ReLU layers, producing the enhanced low-frequency LL' and high-frequency HF' components. LL' serves as the query (Q_f), while HF' is embedded to obtain the keys (K_f) and values (V_f) for a multi-head self attention (MSA) mechanism. The resulting attended features (Att_f) are concatenated with the inverse DWT (IDWT) of the LL' and HF' components and projected to produce the frequency domain feature F_f. The process is expressed as:

$$\begin{cases} (LL', HF') = \text{ReLU}\left(\text{BN}\left(\text{DSC}\left(\text{DWT}\left(F_{in}\right)\right)\right)\right) \\ Q_f = LL', \quad (K_f, V_f) = \text{EMB}(HF') \\ Att_f = \text{MSA}(Q_f, K_f, V_f) \\ F_f = \text{Proj}(Cat(Att_f, \text{IDWT}(LL', HF'))) \end{cases} \tag{13}$$

where $\text{EMB}(\cdot)$ and $\text{Proj}(\cdot)$ represent embedding and projection operations, respectively.

For spatial domain attention, the input F_{in} is first partitioned into non-overlapping windows to reduce computational complexity. The windowed features are linearly projected to obtain the queries (Q_s), keys (K_s), and values (V_s), which are then fed into a windowed MSA mechanism to compute the spatial attended features. These features are concatenated and projected to obtain the spatial domain feature F_s.

Finally, the obtained F_s and F_f are concatenated and projected to yield the final output g, effectively encoding global dependencies of input data. In conjunction with the detail map D_e derived from IAAC, they collectively contribute to capturing optimized detail information with a global perspective.

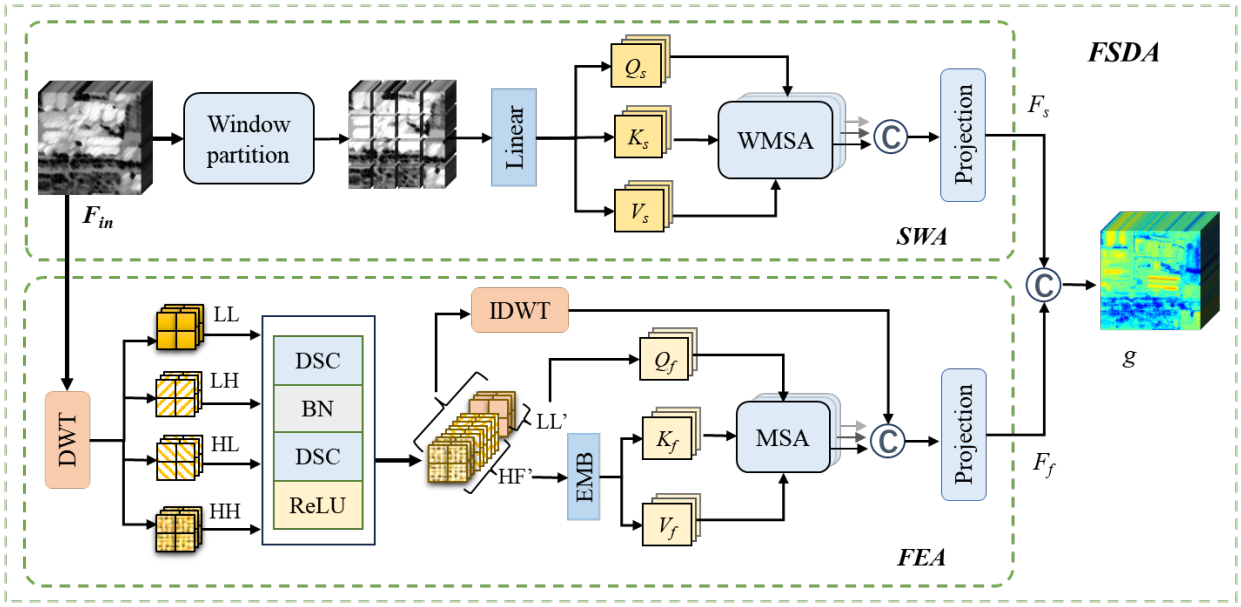


Fig. 6. Flowchart of the FSDA module. DSC, BN, and EMB represent depth-wise separable convolution, batch norm, and embedding, respectively. MSA and WMSA represent multi-head self attention and Windowed MSA respectively.

D. Attention-Integrated Feature Extension Module

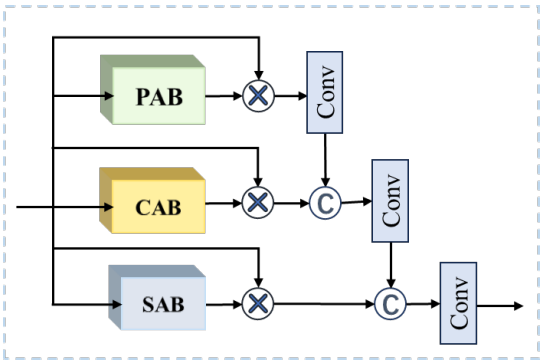


Fig. 7. Flowchart of the attention-integrated feature extension module.

The attention-integrated feature extension module aims to effectively combine multiple attention mechanisms to achieve enhanced feature extension along the channel dimension. This module serves as an intermediate component between the multi-scale residual detail optimization process. As illustrated in Fig. 7, The attention-integrated feature extension module consists of three parallel attention blocks: PAB, CAB, and SAB, similar to those in IAAC. The output of each attention block undergoes element-wise multiplication with the input feature to generate a corresponding attention map. The resulting attention maps are then integrated through a cascaded concatenation process for progressive refinement and integration of features. Specifically, the pixel-attended feature from the PAB is first convolved and then concatenated with the channel-attended feature from the CAB. This concatenated feature is convolved again and subsequently merged with the spatial-attended feature from the SAB through another concatenation operation. This module progressively integrates pixel-level, channel, and spatial-wise attention mechanisms, enabling the expansion of features along the channel dimension while capturing diverse contextual information.

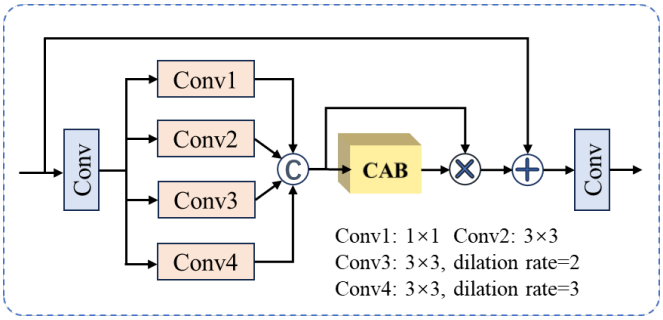


Fig. 8. Flowchart of the reconstruction block

E. Reconstruction Block

The reconstruction block is designed as a residual multi-receptive-field attention mechanism. Specifically, the input features first undergoes a convolutional layer to reduce the channel dimension. To efficiently obtain multi-scale receptive field features, the features are then passed through four parallel convolutional branches, as illustrated in Fig. 8. These four convolutional branches consist of 1×1, 3×3, 3×3 with dilation rate 2, and 3×3 with dilation rate 3 convolutions. The outputs of these four branches are concatenated and then fed into the CAB to extract salient receptive field features. Finally, the reconstruction block incorporates a residual connection, followed by a convolutional layer to match the channel dimension of the output image.

The design of the reconstruction block effectively captures multi-scale receptive field features through parallel convolutional branches with varying dilation rates, while the CAB

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

adaptively emphasizes the most informative receptive field features, thereby yielding high-quality reconstruction results.

F. Loss Function

To preserve spectral-spatial fidelity and maintain perceptual coherence in the fused image, we devise a composite loss function comprising three complementary terms. Specifically, the content loss $\mathcal{L}_c$ serves to maintain the overall spectral and spatial content of the HRMS image, which is formulated as the ℓ_1 norm between the fused output $\boldsymbol{I}_{fused}$ and the GT image $\boldsymbol{I}_{GT}$:

$$\mathcal{L}_c = \left\| \boldsymbol{I}_{fused} - \boldsymbol{I}_{GT} \right\|_1 . \tag{14}$$

To effectively transfer high-frequency spatial information from the PAN image, we introduce a spatial loss $\mathcal{L}_s$ that enforces similarity between the spatial features of the fused output and those of the PAN image, which is defined as:

$$\begin{aligned} \mathcal{L}_s &= \left\| \phi_P(\mathbf{I}_{fused}) - \mathbf{I}_{pan} - (\phi_P(\mathbf{I}_{GT}) - \mathbf{I}_{pan}) \right\|_1 \\ &= \left\| \phi_P(\mathbf{I}_{fused}) - \phi_P(\mathbf{I}_{GT}) \right\|_1, \end{aligned} \tag{15}$$

where ϕ_P denotes a spatial feature extractor that aims to extract features similar to the PAN image.

To enhance perceptual quality and semantic coherence, we employ a perceptual loss that aligns the deep feature representations of the fused output with those of the GT image. leveraging the features extracted from a pre-trained VGG network, the perceptual loss is defined as:

$$\mathcal{L}_p = \left\| \phi_{VGG}(\boldsymbol{I}_{fused}(c1, c2, c3)) - \phi_{VGG}(\boldsymbol{I}_{GT}(c1, c2, c3)) \right\|_1 , \tag{16}$$

where ϕ_{VGG} represents the feature extractor from a pre-trained VGG network, and (c1,c2,c3) represents the first three channels in an MS image.

The total loss function $\mathcal{L}$ is a weighted sum of these three components:

$$\mathcal{L} = \mathcal{L}_c + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_p, \tag{17}$$

where λ_1 and $\lambda_2$are non-negative weights that balance the contributions of each loss term.

IV. EXPERIMENTAL ANALYSIS

A. Experimental Setting

As outlined in Table I, we employed the Pléiades, IKONOS, and WorldView-3 datasets to conduct a extensive assessment of the model's effectiveness. The evaluation included both full-scale (FS) and reduced-scale (RS) experiments, also referred to as real and simulated experiments, respectively. In the RS experiment, we degraded the source images by applying sensor-specific filters tailored to each sensor's modulation transfer function. These degraded images were subsequently downsampled by a factor of four. Adhering to Wald's protocol [5], the original MS images served as the ground truth (GT) references. For the FS experiment, we performed image fusion at the original scale due to the unavailability of GT images.

In the experiments, each satellite dataset comprised 2,500 training samples. To augment each dataset, we employed random cropping during the training process. The batch size was set to 25 to align with the hardware capabilities. The training spanned 300 epochs with an initial learning rate of 0.0006, which was halved every 100 epochs to accommodate the diminishing gradients. Besides, an independent test set of 100 image groups was held out for evaluation purposes. The hyperparameters λ_1 and λ_2 in Equation (17) are empirically set to 1 and 0.1, respectively, to balance the order of magnitude across the loss terms. All the experiments were performed on a computer featuring an RTX-3090 GPU.

To validate the effectiveness of our method, we conduct comprehensive experiments comparing our method against a series of traditional and state-of-the-art pansharpening techniques. For classical methods, we consider GSA [7] and AWLP-R [10] algorithms. Among CNN-based approaches, we evaluate APNN-FT [36], VO+Net [24], FusionNet [23], PCGAN [20], and TDNet [29], the transformer-based PMRF [28], and DCPNet [21] in our comparisons. For fairness, all deep learning models were retrained on our datasets. Furthermore, we incorporate the EXP method as a spectral benchmark through upsampling, though it does not participate in the pansharpening comparison.

We evaluate the performance using widely adopted objective quality metrics. For RS experiments, UIQI↑, Q2n↑, ERGAS↓, SAM↓, and SCC↑ [3], [37] are employed to assess both spectral and spatial fidelities. For FS experiments, the spectral distortion metric D_λ↓, the spatial distortion metric D_s↓, and the overall quality metric QNR↑, which is derived from D_λ and D_s [38] are employed. An upward arrow (↑) indicates that higher values represent better performance, while a downward arrow (↓) denotes that lower values are preferable.

B. Reduced-scale Experiments

Fig. 9 showcases the pansharpening results for remote sensing images from the IKONOS dataset, displaying only the RGB channels for visual effect. From the zoomed-in yellow bounding boxes, it can be observed that the outputs of PMRF, FusionNet, and TDNet exhibit noticeable blurring artifacts. In contrast, GSA and VO+Net tend to over-enhance spatial details, leading to spectral distortions in forest regions. PCGAN's result appears overly dark, suffering from severe spectral distortions. Both DCPNet and our method produce fusion outputs that are visually consistent with the ground truth (GT) image. This observation is further corroborated by the absolute error maps (AEMs) displayed in Fig. 10, where our approach exhibits the least residual errors.

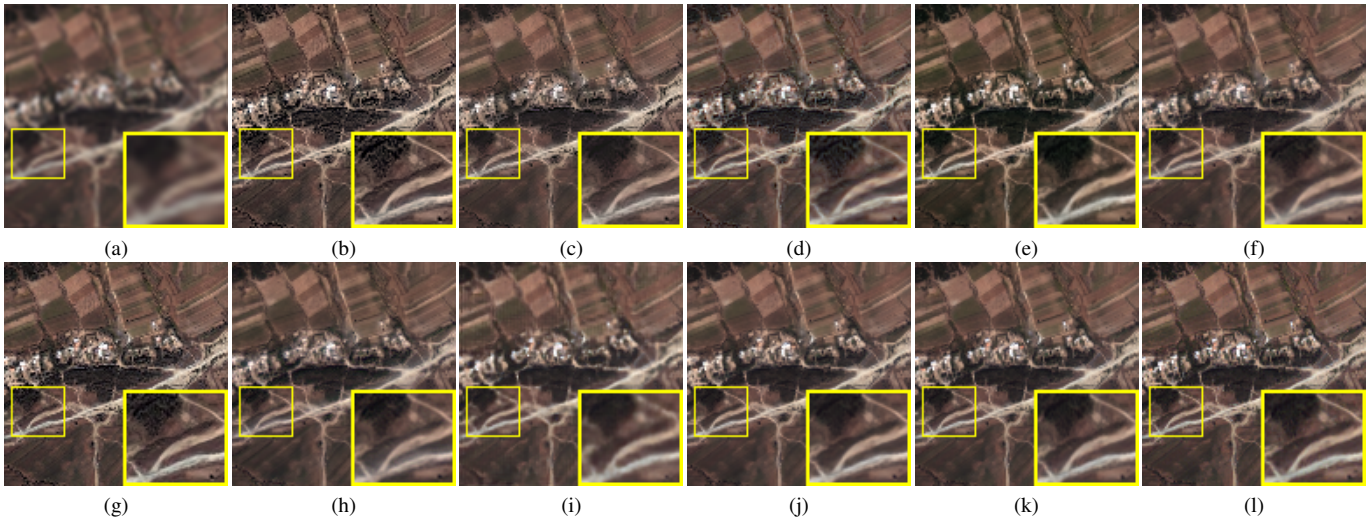


Fig. 9. Pansharpened RS images in IKONOS dataset. (a) EXP. (b) GSA. (c) AWLP-R. (d) APNN-FT. (e) PCGAN. (f) FusionNet. (g) VO + Net. (h) TDNet. (i) PMRF. (j) DCPNet. (k) Proposed. (l) GT.

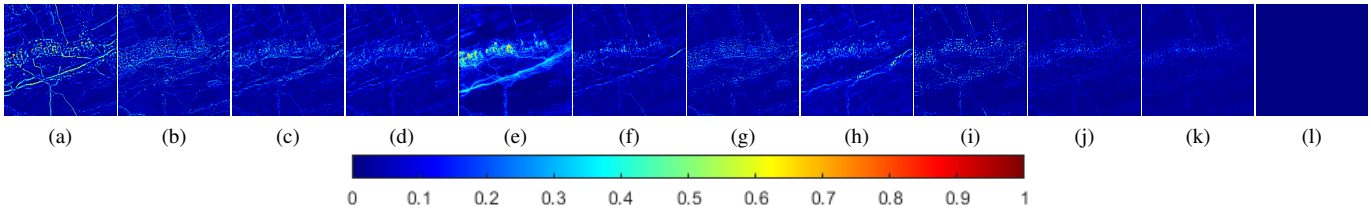


Fig. 10. AEMs of the fusion results depicted in Fig. 9. (a) EXP. (b) GSA. (c) AWLP-R. (d) APNN-FT. (e) PCGAN. (f) FusionNet. (g) VO + Net. (h) TDNet. (i) PMRF. (j) DCPNet. (k) Proposed. (l) GT.

TABLE II
OBJECTIVE ASSESSMENT OF FUSION RESULTS PRESENTED IN FIG. 9 AND MEAN ASSESSMENT ON IKONOS DATASET

Methods	Fig. 9					Mean				
	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑
EXP [39]	0.7754	3.4592	4.4058	0.7015	0.7743	0.7627	3.3579	4.0487	0.6849	0.7574
GSA [7]	0.8762	5.5192	3.8820	0.8489	0.8662	0.8438	4.5543	3.9428	0.8284	0.8351
AWLP-R [10]	0.9074	3.8645	3.1646	0.8886	0.9032	0.8682	3.5731	3.3621	0.8540	0.8633
APPN-FT [36]	0.9039	4.5928	2.9787	0.8733	0.9030	0.8780	4.2710	3.3817	0.8449	0.8673
PCGAN [20]	0.9165	4.6391	3.0596	0.9596	0.9185	0.8412	8.4360	6.0984	0.9463	0.8224
FusionNet [23]	0.9434	3.1014	2.3992	0.9252	0.9399	0.9118	3.1394	2.6173	0.8973	0.9088
VO+Net [24]	0.9230	4.1683	3.0574	0.8873	0.9165	0.8944	3.5710	3.1333	0.8685	0.8866
TDNet [29]	0.9108	4.2215	3.0303	0.9083	0.9034	0.8877	3.7885	3.0256	0.8896	0.8783
PMRF [28]	0.9282	3.1649	3.1144	0.8925	0.8789	0.8984	3.2853	3.5434	0.8643	0.8381
DCPNet [21]	0.9586	2.6763	1.9877	0.9422	0.9549	0.9280	2.7920	2.2970	0.9161	0.9242
Proposed	**0.9684**	**2.3973**	**1.6826**	**0.9606**	**0.9657**	**0.9539**	**2.3312**	**1.6918**	**0.9507**	**0.9520**

The objective quality metrics for the example in Fig. 9, along with the mean metrics across all test images on IKONOS dataset, are reported in Table II. The best-performing values are highlighted in bold. It can be seen that our method achieves the optimal scores across all evaluated metrics, demonstrating a clear advantage and validating the effectiveness of our proposed approach.

The pansharpening results for remote sensing images from the Pléiades dataset are illustrated in Fig. 11. From the zoomed-in yellow bounding boxes, it can be observed that the outputs of PCGAN and TDNet suffer from severe spectral distortions, where the red rooftops are erroneously rendered in orange hues. The results of PMRF, APNN-FT, and FusionNet exhibit blurring artifacts. In contrast, GSA, AWLP-R, and VO+Net tend to over-enhance spatial details, leading to minor spectral distortions. Our proposed method yields a fusion output that visually aligns most closely with the ground truth (GT) image. This observation is further corroborated by the absolute error maps (AEMs) displayed in Fig. 12, where our approach exhibits the least residual errors.

The quality metrics for the example in Fig. 11 and the average metrics across all test are shown in Table III. It can be observed that our method achieves the optimal performance across all evaluated metrics, demonstrating its effectiveness.

The pansharpening results for images from the WorldView-3 dataset are presented in Fig. 13. From the zoomed-in

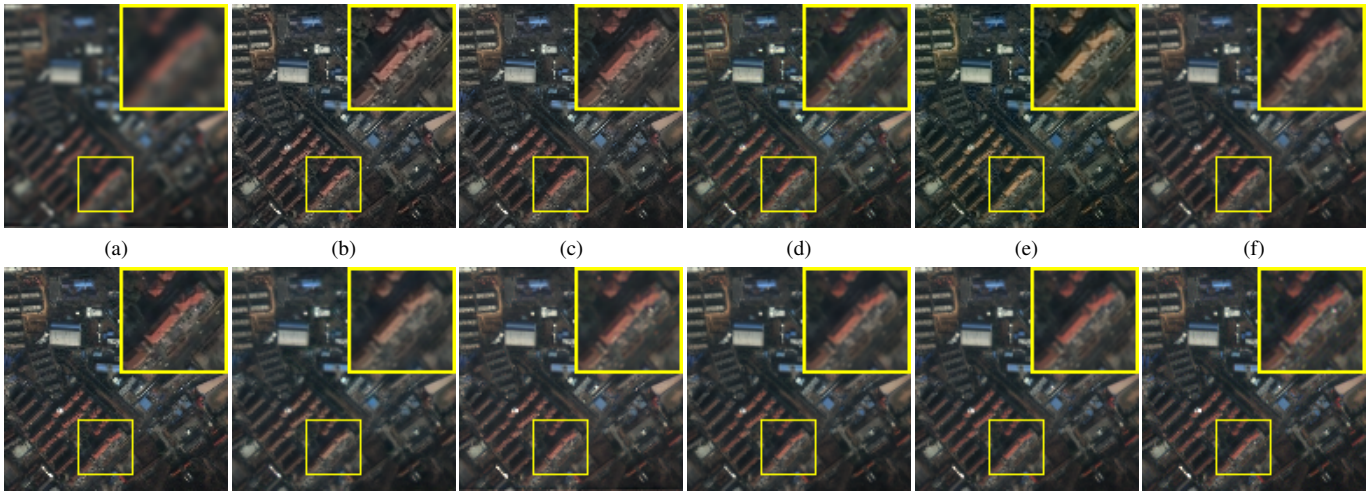


Fig. 11. Pansharpened RS images in Pléiades dataset. (a) EXP. (b) GSA. (c) AWLP-R. (d) APNN-FT. (e) PCGAN. (f) FusionNet. (g) VO + Net. (h) TDNet. (i) PMRF. (j) DCPNet. (k) Proposed. (l) GT.

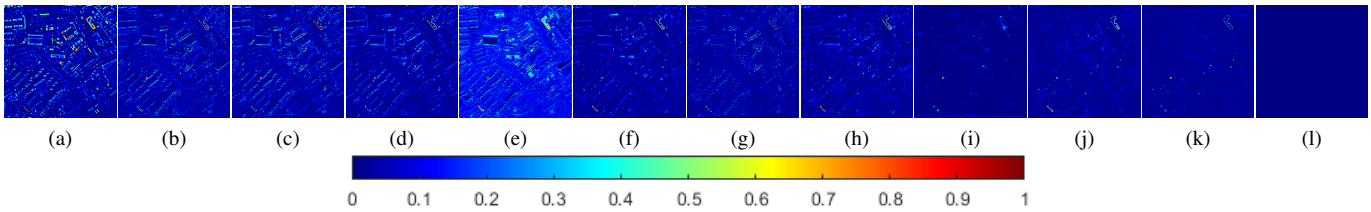


Fig. 12. AEMs of the fusion results depicted in Fig. 11. (a) EXP. (b) GSA. (c) AWLP-R. (d) APNN-FT. (e) PCGAN. (f) FusionNet. (g) VO + Net. (h) TDNet. (i) PMRF. (j) DCPNet. (k) Proposed. (l) GT.

TABLE III
OBJECTIVE ASSESSMENT OF FUSION RESULTS PRESENTED IN FIG. 11 AND MEAN ASSESSMENT ON PLÉIADES DATASET

Methods	Fig. 11					Mean				
	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑
EXP [39]	0.8341	3.6853	5.1986	0.7481	0.8247	0.8194	3.0838	3.9728	0.7522	0.8173
GSA [7]	0.9010	4.2086	4.3586	0.8420	0.8943	0.8751	3.3739	3.5754	0.8296	0.8709
AWLP-R [10]	0.9030	3.8760	4.3814	0.8431	0.8997	0.8863	3.0752	3.4263	0.8332	0.8846
APPN-FT [36]	0.9300	4.3748	3.6785	0.8743	0.9275	0.9055	3.6274	3.0794	0.8533	0.9014
PCGAN [20]	0.8977	8.8853	5.5755	0.8644	0.8605	0.8653	7.7779	5.0045	0.8483	0.8052
FusionNet [23]	0.9557	3.3087	2.8635	0.9261	0.9524	0.9282	2.8778	2.9819	0.9039	0.9262
VO+Net [24]	0.9438	3.1869	3.2358	0.8890	0.9404	0.9122	2.7449	2.9123	0.8660	0.9090
TDNet [29]	0.9445	4.3676	3.2233	0.9224	0.9398	0.9215	3.5946	2.6642	0.9095	0.9164
PMRF [28]	0.9710	3.1459	2.9423	0.9506	0.9480	0.9508	2.8368	2.6512	0.9451	0.9122
DCPNet [21]	0.9759	2.9138	2.1270	0.9647	0.9732	0.9528	2.5253	1.7756	0.9476	0.9520
Proposed	**0.9842**	**2.5627**	**1.6079**	**0.9733**	**0.9821**	**0.9667**	**2.2254**	**1.3415**	**0.9682**	**0.9667**

yellow boxes, we can observe that the output of TDNet exhibits noticeable spectral distortions, since the rooftops, which should appear orange, have changed color in their results. The result of APNN-FT suffers from blurring artifacts. GSA, AWLP-R, VO+Net, and DCPNet tend to over-enhance edge details, leading to halo effects. Our proposed method yields a fusion output that visually aligns most closely with GT. This observation is further substantiated by the corresponding AEMs, where our approach exhibits the least residual errors. The objective quality metrics for the fusion results from the WorldView-3 dataset, are reported in Table IV. It can be seen that, similar to the results of other datasets, our method achieves the optimal performance across all evaluated metrics.

C. Full-scale Experiments

The FS experiment performs real image fusion with the fused image size of $1024 \times 1024 \times B$, where B denotes the number of spectral bands. Since the real image dimensions are excessively large, this work presents cropped sections of the fused images for display purposes. The results of FS experiments for the IKONOS dataset are illustrated in Fig. 15. It can be observed that although the GSA method yields sharp spatial details, it tends to over-enhance the high-frequency components, causing spectral distortions in the red rooftops. The outputs of APNN-FT and VO+Net exhibit spatial distortions, with distorted edge structures. PCGAN, FusionNet, and TDNet suffer from severe spectral distortions, as evident

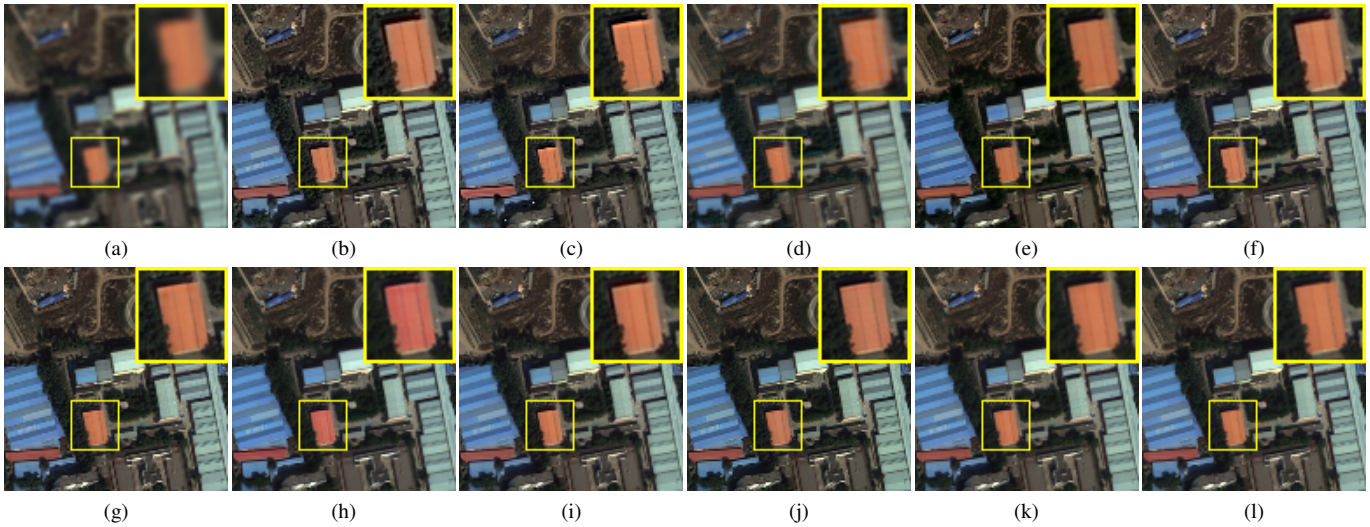


Fig. 13. Pansharpened RS images in WorldView-3 dataset. (a) EXP. (b) GSA. (c) AWLP-R. (d) APNN-FT. (e) PCGAN. (f) FusionNet. (g) VO + Net. (h) TDNet. (i) PMRF. (j) DCPNet. (k) Proposed. (l) GT.

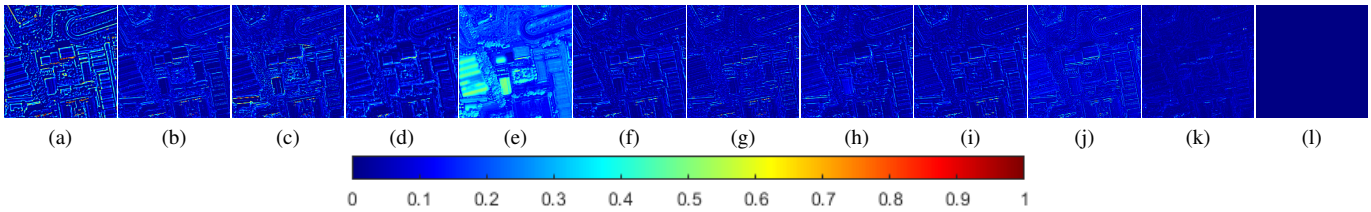


Fig. 14. AEMs of the fusion results depicted in Fig. 13. (a) EXP. (b) GSA. (c) AWLP-R. (d) APNN-FT. (e) PCGAN. (f) FusionNet. (g) VO + Net. (h) TDNet. (i) PMRF. (j) DCPNet. (k) Proposed. (l) GT.

TABLE IV
OBJECTIVE ASSESSMENT OF FUSION RESULTS PRESENTED IN FIG. 13 AND MEAN ASSESSMENT ON WORLDVIEW-3 DATASET

Methods	Fig. 13					Mean				
	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q8↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q8↑
EXP [39]	0.8390	4.0151	4.9580	0.7289	0.8228	0.6567	5.4430	6.1138	0.6284	0.6553
GSA [7]	0.9296	6.2090	3.4790	0.8583	0.9373	0.8589	6.4646	4.2272	0.8384	0.8720
AWLP-R [10]	0.9358	4.7447	3.6970	0.8707	0.9316	0.8769	5.4589	4.0417	0.8599	0.8805
APPN-FT [36]	0.9359	5.5699	3.3232	0.8995	0.9365	0.8065	6.7477	4.7598	0.8091	0.8208
PCGAN [20]	0.9415	6.3988	3.9073	0.9476	0.9519	0.8312	8.3228	5.6929	0.8971	0.8302
FusionNet [23]	0.9640	3.8646	2.6493	0.9120	0.9597	0.8903	5.1457	3.7813	0.8736	0.8936
VO+Net [24]	0.9548	4.8212	2.9154	0.8923	0.9522	0.8842	5.7228	3.8555	0.8657	0.8846
TDNet [29]	0.9652	3.9940	2.6844	0.9178	0.9614	0.8912	5.3467	3.7361	0.8787	0.8936
PMRF [28]	0.9661	3.9695	2.5746	0.9163	0.9620	0.8968	5.1627	3.6081	0.8797	0.9005
DCPNet [21]	0.9717	3.5783	2.5516	0.9473	0.9622	0.9234	4.8348	3.0973	0.9288	0.9187
Proposed	**0.9855**	**2.9720**	**1.5926**	**0.9682**	**0.9834**	**0.9363**	**4.3225**	**2.5876**	**0.9472**	**0.9387**

from the color shifts in the rooftops. The PMRF result appears blurred and lacks adequate spatial enhancement. In contrast, the DCPNet method delivers a relatively smooth output but falls short in spectral enhancement.

Our approach strikes an effective balance between spatial detail enhancement and spectral fidelity preservation, as evidenced by the sharper spatial and more faithful color representation in the fused outputs. Table V further reports the objective quality metrics for the example in Fig. 15 and the mean metrics across the IKONOS test set. It can be seen that our method achieves optimal performance in terms of the D_s and QNR metrics, while ranking second for the D_λ metric. These quantitative results further validate the effectiveness of our approach in preserving spectral and spatial quality across multiple scales.

D. Ablation Study

1) The Impact of the Proposed Modules: In this section, an ablation study is conducted to evaluate the effectiveness of the proposed modules. Our model mainly includes the IAAC module, the FSDA module comprising FEA and SWA, the attention-integrated feature extension (AFE) module, and the residual multi-receptive fields reconstruction (RMFR) block. Thus, we construct the ablation models as follows:

Baseline: This model uses an multiscale and multi-stage architecture similar to our proposed model, but replaces the

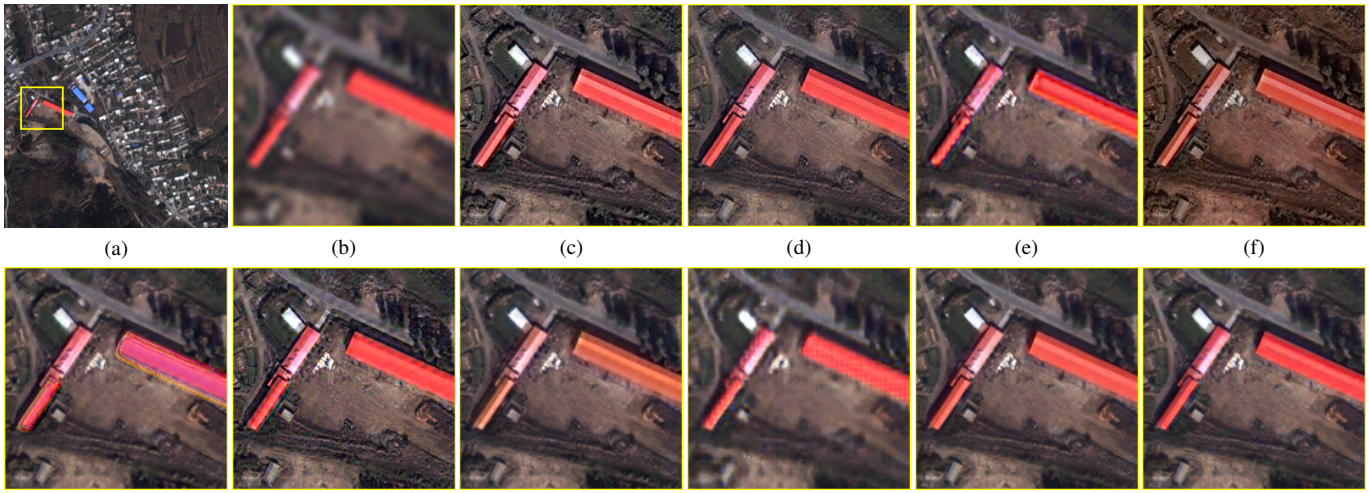


Fig. 15. Pansharpened FS images in IKONOS dataset. (a) UPMS. (b) EXP. (c) GSA (d) AWLP-R. (e) APNN-FT. (f) PCGAN. (g) FusionNet. (h) VO + Net. (i) TDNet. (j) PMRF. (k) DCPNet. (l) Proposed.

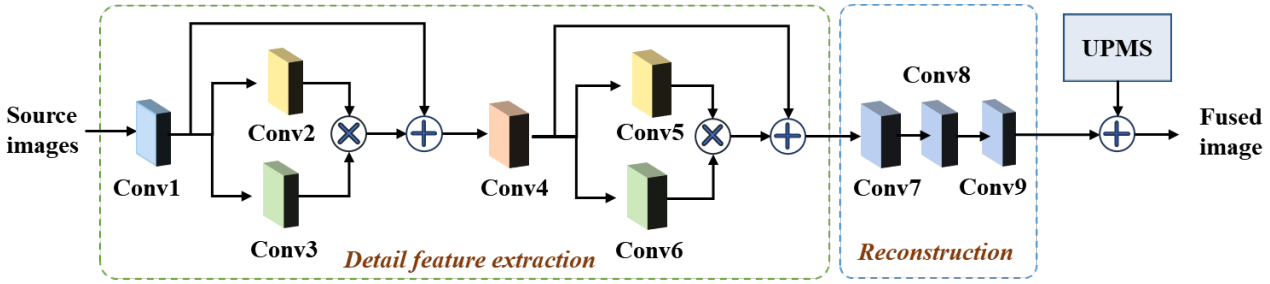


Fig. 16. The flowchart of the baseline model.

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

Methods	Fig. 15			Mean		
	$D_\lambda\downarrow$	$D_s\downarrow$	QNR↑	$D_\lambda\downarrow$	$D_s\downarrow$	QNR↑
EXP [39]	0.0005	0.2254	0.7743	0.0006	0.2334	0.7662
GSA [7]	0.1009	0.1311	0.7682	0.1026	0.1406	0.7574
AWLP-R [10]	0.1536	0.1399	0.7280	0.1456	0.1597	0.7195
APPN-FT [36]	0.0693	0.0378	0.8943	0.0624	0.0639	0.8784
PCGAN [20]	0.1411	0.1966	0.6901	0.1305	0.2374	0.6676
FusionNet [23]	0.0894	0.0675	0.8491	0.0772	0.0663	0.8618
VO+Net [24]	0.1383	0.1162	0.7615	0.1246	0.1219	0.7713
TDNet [29]	0.1190	0.0778	0.8125	0.1044	0.0772	0.8267
PMRF [28]	**0.0131**	0.1461	0.8427	**0.0236**	0.1567	0.8233
DCPNet [21]	0.0786	0.0489	0.8764	0.0754	0.0557	0.8735
Proposed	0.0686	**0.0287**	**0.9047**	0.0350	**0.0628**	**0.9022**

proposed modules with pure convolution operations, consisting of Conv1 to Conv9, as shown in Fig. 16.

IAAC model: This model replaces Conv2 and Conv5 in the baseline model with the IAAC module to investigate its impact on performance.

IAAC + FEA model: Building on the IAAC model, this variant further replaces Conv3 and Conv6 with the FEA module.

IAAC + SWA model: Similar to the IAAC + FEA model, this variant replaces the FEA module with the SWA module.

IAAC + FSDA model: Combining both FEA and SWA, this model incorporates the FSDA module, replacing the FEA module in the IAAC + FEA model.

IAAC + FSDA + AFE model: This model enhances the IAAC + FSDA model by adding the AFE module.

The proposed model: This model combines all the proposed modules (IAAC, FSDA, AFE, and RMFR) with the baseline to evaluate the overall performance improvement achieved by the complete model.

Using an image pair from the IKONOS dataset as an example, the subjective evaluation is shown in Fig. 17. From the enlarged yellow box and the corresponding AEMs, we can observe that the baseline model performs worse compared to other ablation models. With the inclusion of the proposed modules, the fusion quality improves, indicating that these modules can progressively correct the details. The complete proposed model, incorporating all modules, achieves the best performance and closely aligns with the GT. This observation is further supported by the average objective evaluation results presented in Table VI, which demonstrate the effectiveness of the proposed components.

2) Visualization of Intermediate Results: To further illustrate the impact of the proposed components on detail correction, we visualize the intermediate features using the image shown in Fig. 17. The results are presented in Fig. 18. As shown in Fig. 18(a), the IAAC module effectively

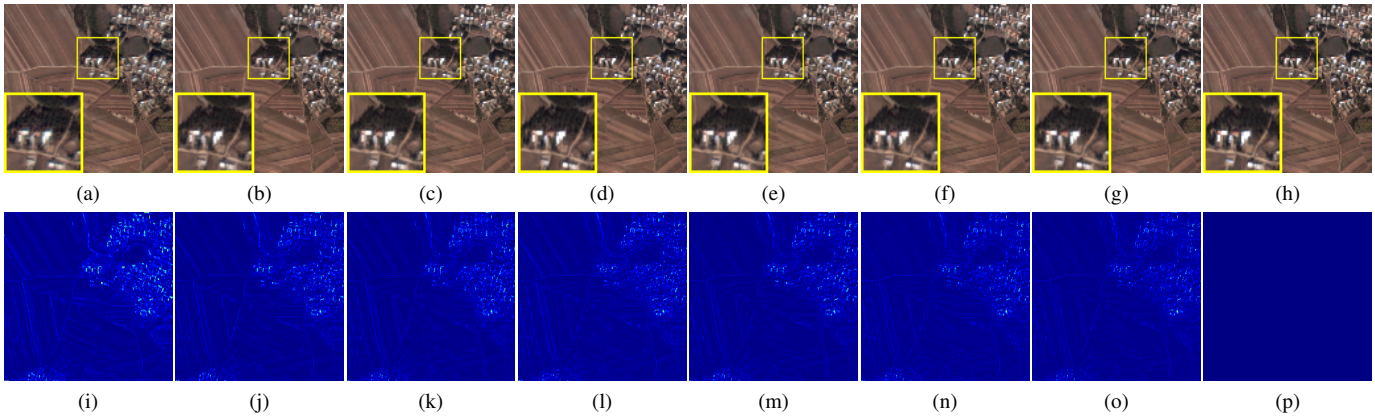


Fig. 17. Pansharpening results on IKONOS dataset using different ablation models. (a) Baseline. (b) IAAC. (c) IAAC + FEA. (d) IAAC + SWA. (e) IAAC + FSDA. (f) IAAC + FSDA + AFE. (g) Proposed. (h) GT. (i)-(p) is the corresponding AEMs of (a)-(h).

TABLE VI
AVERAGE OBJECTIVE EVALUATION OF ABLATION MODELS ON THE IKONOS DATASET

Models	IAAC	FEA	SWA	AFE	RMFR	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑
Baseline	✗	✗	✗	✗	✗	0.9029	3.3936	2.7865	0.8781	0.8928
IAAC	✓	✗	✗	✗	✗	0.9361	2.7083	2.1119	0.9240	0.9316
IAAC+FEA	✓	✓	✗	✗	✗	0.9417	2.6611	2.0176	0.9280	0.9393
IAAC+SWA	✓	✗	✓	✗	✗	0.9409	2.6889	2.0123	0.9302	0.9380
IAAC+FSDA	✓	✓	✓	✗	✗	0.9482	2.4529	1.8389	0.9414	0.9461
IAAC+FSDA+AFE	✓	✓	✓	✓	✗	0.9505	2.3991	1.7627	0.9470	0.9481
Proposed	✓	✓	✓	✓	✓	**0.9539**	**2.3312**	**1.6918**	**0.9507**	**0.9520**

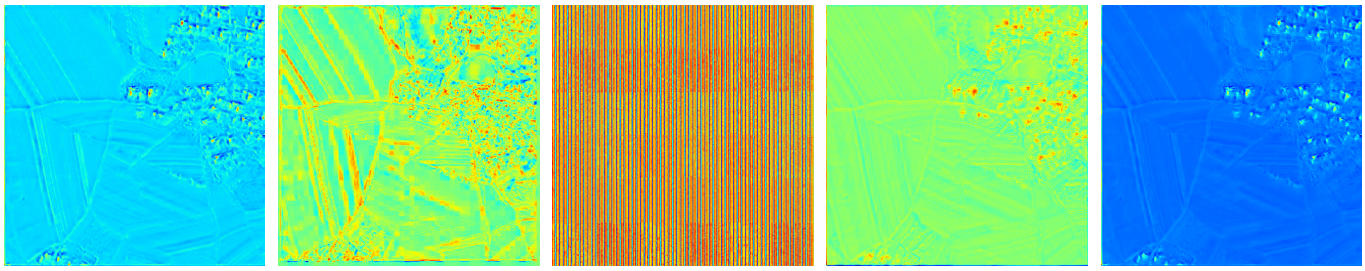


Fig. 18. Visualization of intermediate results obtained by: (a) IAAC, (b) SWA, (c) FEA, (d) FSDA, and (e) combining IAAC and FSDA.

extracts rich initial local features. The SWA module captures long-range spatial contextual features through window-wise spatial attention, as seen in Fig. 18(b). With the introduction of frequency domain attention, the FEA module extracts globally relevant features, as demonstrated in Fig. 18(c). When SWA and FEA are integrated into the FSDA module, the resulting features (Fig. 18(d)) exhibit more prominent and useful global features while suppressing irrelevant ones, further enhancing the local features extracted by IAAC. The final integrated feature, presented in Fig. 18(e), illustrates that the local-global joint optimization mechanism leads to more distinct feature representations compared to the features derived from either IAAC or FSDA alone.

3) The Impact of the Loss Function: In this work, three loss function terms are defined, as shown in Equation (17), including $\mathcal{L}_c$, $\mathcal{L}_s$, and $\mathcal{L}_p$. A series of ablation experiments were conducted to validate the effectiveness of these three terms and the impact of their weighting parameters λ_1 and λ_2. For the fusion results of the RS experiments exhibit minor differences, making it difficult to distinguish, we primarily present the objective evaluation from the FS experiments (i.e., real experiments), where the differences are more significant.

Since $\mathcal{L}_c$ is crucial for maintaining content similarity with the GT, we evaluated the fusion results obtained using solely the $\mathcal{L}_c$ loss term, as well as combinations of multiple loss terms. For parameter settings, we first determine the optimal λ_1 by searching for the best QNR value. Once λ_1 is fixed, we then identify the optimal λ_2. Using the IKONOS dataset as a representative example, the average metrics are summarized in Table VII. From the table, it can be observed that when only $\mathcal{L}_c$ is employed (i.e., $\lambda_1 = 0$), both spectral and spatial qualities are relatively suboptimal. Introducing $\mathcal{L}_s$, which focuses on preserving spatial details, leads to a significant improvement in spatial quality. When $\lambda_1 = 1$, the optimal balance between $\mathcal{L}_c$ and $\mathcal{L}_s$ is achieved. Subsequently, by fixing λ_1 and incorporating $\mathcal{L}_p$, the spectral quality is further

TABLE VII
AVERAGE OBJECTIVE EVALUATION FOR DIFFERENT COMBINATIONS OF LOSS FUNCTION TERMS

Parameters	values	$D_\lambda\downarrow$	$D_s\downarrow$	QNR↑
λ_1	0.0	0.0416	0.0807	0.8808
	0.5	0.0473	0.0702	0.8862
	1.0	0.0603	0.0538	**0.8904**
	1.5	0.0640	0.0501	0.8883
λ_2	0.00	0.0603	0.0538	0.8904
	0.05	0.0497	0.0598	0.8942
	0.10	0.0350	0.0628	**0.9022**
	0.15	0.0316	0.0685	0.8953

TABLE VIII
EFFICIENCY COMPARISON. NOP REPRESENTS NUMBER OF PARAMETERS.

Methods	Testing Time(s)	Model Size(M)	FLOPs(G)	NoP(M)
EXP [39]	-	-	-	-
GSA [7]	0.035(CPU)	-	-	-
AWLP-R [10]	0.042(CPU)	-	-	-
APPN-FT [36]	3.232(CPU)	-	-	0.31
PCGAN [20]	0.043(GPU)	124.15	11.16	32.62
FusionNet [23]	0.001(GPU)	0.29	9.92	0.23
VO+Net [24]	11.652(CPU)	-	-	0.31
TDNet [29]	0.003(GPU)	2.15	19.98	0.49
PMRF [28]	0.004(GPU)	1.60	32.62	0.39
DCPNet [21]	0.238(GPU)	8.27	229.56	1.96
Proposed	0.003(GPU)	0.63	6.46	0.11

enhanced. The combination of $\mathcal{L}_c$, $\mathcal{L}_s$, and $\mathcal{L}_p$ achieves the best QNR metric when $\lambda_1 = 1$ and $\lambda_2 = 0.1$.

E. Efficiency Analysis

To assess the computational efficiency of our method, we conducted a comprehensive analysis by measuring various performance metrics. Specifically, we assessed the testing time, the model weight size after training, FLOPs, NoP required by our method and other SOTA deep learning-based approaches for processing a single input sample. The input sample consisted of two tensors with dimensions of $1 \times 1 \times 256 \times 256$ and $1 \times 4 \times 64 \times 64$, respectively. To ensure a fair comparison, all methods were evaluated on the same hardware platform.

The comparative results are summarized in Table VIII. Our proposed method demonstrated superior computational efficiency, achieving the lowest FLOPs, fewest NoP, and the second smallest model size, compared to all other deep learning-based methods included in the evaluation. Notably, the testing time on GPU was comparable across different deep learning-based approaches, benefiting from hardware acceleration.

The exceptional efficiency of our method stems from the concise and interpretable design of IAAC and FSDA, built upon the new detail injection algorithm, along with the use of multi-receptive field dilated convolutions in the reconstruction block. These architectural components effectively reduce computational complexity while maintaining high performance, enabling efficient deployment and execution on various hardware platforms.

TABLE IX
OBJECTIVE ASSESSMENT OF THE CLASSIFICATION RESULTS IN FIG.19

Methods	OA↑	KC↑
EXP [39]	0.6795	0.5762
GSA [7]	0.7286	0.6392
AWLP-R [10]	0.7644	0.6875
APPN-FT [36]	0.8034	0.7403
PCGAN [20]	0.5169	0.3640
FusionNet [23]	0.8458	0.7955
VO+Net [24]	0.7933	0.7253
TDNet [29]	0.8036	0.7396
PMRF [28]	0.7741	0.7005
DCPNet [21]	0.8543	0.8208
Proposed	**0.8932**	**0.8583**

F. Downstream Application

To validate the effectiveness of our fusion results for downstream tasks, we performed scene classification and subsequent segmentation using the ENVI tool. Taking the fusion result in Fig. 9 as an example, the scene classification and segmentation results are shown in Fig. 19. As evident from the figure, the fusion results obtained by methods such as GSA, AWLP-R, PCGAN, and PMRF exhibit suboptimal performance in the classification task, deviating significantly from the GT. In contrast, our fusion result demonstrates the closest resemblance to the GT in terms of classification performance, outperforming the other methods.

To further objectively evaluate the classification accuracy, we employed two representative metrics, including the kappa coefficient (KC↑), overall accuracy (OA↑). The corresponding objective metrics for Fig. 19 are presented in Table IX. The table also demonstrates the superior accuracy of our classification results, corroborating the effectiveness of our proposed method for downstream applications.

G. More Discussion

The proposed method integrates an invertible attention guided adaptive convolution module and dual-domain attention mechanism within a multiscale residual structure, which collectively aim to address the challenges of feature misalignment and efficient detail extraction in HRMS image reconstruction. The proposed invertible attention mechanism focuses on extracting spatial-spectral-aware detail features from local regions efficiently and losslessly. By embedding spectral-spatial attention, the model dynamically integrates local features while preserving critical information, enabling it to address small-scale misalignments.

To complement the local feature correction, the frequency-spatial dual-domain attention mechanism captures long-range dependencies in both frequency and spatial domains. The frequency-enhanced Transformer extracts global contextual information in the frequency domain, correcting misaligned features that span across larger areas. The spatial window Transformer operates on window-wise spatial features with a fine-grained focus, ensuring precise alignment at the window scale.

The integration of local detail features and long-range dependencies ensures that the model can simultaneously cor-

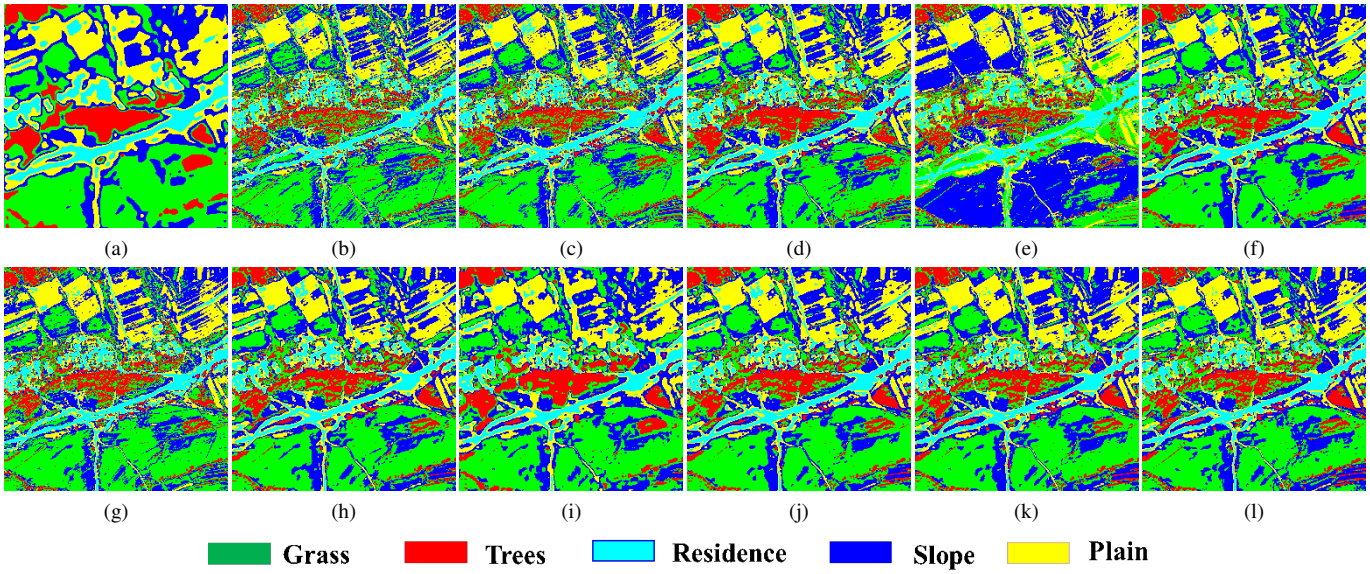


Fig. 19. Classification and segmentation results of the fused images in Fig. 9. (a) EXP. (b) GSA. (c) AWLP-R. (d) APNN-FT. (e) PCGAN. (f) FusionNet. (g) VO + Net. (h) TDNet. (i) PMRF. (j) DCPNet. (k) Proposed. (l) GT.

rect small-scale local misalignments and global contextual inconsistencies. This is further enhanced by the residual multi-receptive field attention block in the reconstruction stage, which combines multi-scale features for a well-aligned and high-quality HRMS image.

Our method can also be extended to other image processing tasks beyond HRMS reconstruction, such as hyperspectral image fusion, super-resolution, and cross-domain image synthesis. The modularity of the adaptive convolution and attention mechanisms allows for easy integration into other deep learning frameworks, potentially benefiting a wide range of computer vision and geospatial analysis applications.

Despite its advantages, the proposed method has some limitations. The current model relies heavily on the quality of the input data. In scenarios where the input images are heavily degraded or contain significant noise, the performance of the model may degrade. Future work could explore incorporating noise-robust mechanisms or pre-processing techniques to enhance the method's applicability in such challenging scenarios.

V. Conclusion

To address the limitations of current pansharpening methods in terms of local detail misalignments, interpretability, and efficiency, we propose a novel detail-injection algorithm that incorporates joint correction of local and global details. Based on this algorithm, we further develop a pansharpening network, termed IACDT. This network incorporates an invertible adaptive convolution module guided by spatial and spectral attention, thereby efficiently enhancing the extraction of local detailed information. Additionally, we introduce a dual-domain attention mechanism that synergistically combines a frequency enhanced transformer with a spatial window transformer to facilitate global feature correction. The adaptive convolution module and dual-domain attention jointly optimize detail extraction within a multiscale residual structure, while the HRMS image is generated through a residual multi-receptive field attention mechanism. Furthermore, we meticulously design a compound loss function, which includes content loss, spatial loss, and perceptual loss, to guide the generation of fused images that maintain spatial details and spectral fidelity. Extensive experiments demonstrate the efficacy of our approach, highlighting its superior fusion performance and efficiency relative to other state-of-the-art methods. Additionally, our method exhibits enhanced performance in downstream tasks, further corroborating its practical utility and theoretical robustness.

References

[1] G. Vivone, L.-J. Deng, S. Deng, D. Hong, M. Jiang, C. Li, W. Li, H. Shen, X. Wu, J.-L. Xiao, J. Yao, M. Zhang, J. Chanussot, S. García, and A. Plaza, "Deep Learning in Remote Sensing Image Fusion: Methods, protocols, data, and future perspectives," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–43, 2024.

[2] M. Ciotola, G. Guarino, G. Vivone, G. Poggi, J. Chanussot, A. Plaza, and G. Scarpa, "Hyperspectral Pansharpening: Critical review, tools, and future perspectives," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–29, 2024.

[3] 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.

[4] T. Xu, T.-Z. Huang, L.-J. Deng, J.-L. Xiao, C. Broni-Bediako, J. Xia, and N. Yokoya, "A Coupled Tensor Double-Factor Method for Hyperspectral and Multispectral Image Fusion," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–17, 2024.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] N. H. Kaplan, I. Erer, O. Ozcan, and N. Musaoglu, "MTF driven adaptive multiscale bilateral filtering for pansharpening," *International Journal of Remote Sensing*, vol. 40, no. 16, pp. 6262–6282, Aug. 2019.

[13] 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.

[14] J.-L. Xiao, T.-Z. Huang, L.-J. Deng, Z.-C. Wu, X. Wu, and G. Vivone, "Variational Pansharpening Based on Coefficient Estimation With Non-local Regression," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 110–125, 2023.

[15] 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.

[16] Y. Yang, H. Lu, S. Huang, Y. Fang, and W. Tu, "An efficient and high-quality pansharpening model based on conditional random fields," *Inf. Sci.*, vol. 553, pp. 1–18, Apr. 2021.

[17] R. Wen, L.-J. Deng, Z.-C. Wu, X. Wu, and G. Vivone, "A Novel Spatial Fidelity With Learnable Nonlinear Mapping for Panchromatic Sharpening," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–15, 2023.

[18] 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.

[19] Z. Su, Y. Yang, S. Huang, W. Wan, W. Tu, H. Lu, and C. Chen, "CTCP: Cross transformer and CNN for pansharpening," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 3003–3011.

[20] 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.

[21] Y. Zhang, X. Yang, H. Li, M. Xie, and Z. Yu, "DCPNet: A Dual-Task Collaborative Promotion Network for Pansharpening," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–16, 2024.

[22] J. Li, K. Zheng, L. Gao, L. Ni, M. Huang, and J. Chanussot, "Model-Informed Multistage Unsupervised Network for Hyperspectral Image Super-Resolution," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–17, 2024.

[23] 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.

[24] 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.

[25] P. Wang, Z. He, B. Huang, M. D. Mura, H. Leung, and J. Chanussot, "VOGTNet: Variational Optimization-Guided Two-Stage Network for Multispectral and Panchromatic Image Fusion," *IEEE Trans. Neural Netw. Learning Syst.*, pp. 1–15, 2024.

[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] K. Zhang, Z. Li, F. Zhang, W. Wan, and J. Sun, "Pan-Sharpening Based on Transformer With Redundancy Reduction," *IEEE Geosci. Remote Sensing Lett.*, vol. 19, pp. 1–5, 2022.

[28] 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.

[29] 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.

[30] J. Cheng, H. Liu, T. Liu, F. Wang, and H. Li, "Remote sensing image fusion via wavelet transform and sparse representation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 104, pp. 158–173, 2015.

[31] W. Ma, Z. Pan, J. Guo, and B. Lei, "Achieving Super-Resolution Remote Sensing Images via the Wavelet Transform Combined With the Recursive Res-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3512–3527, Jun. 2019.

[32] S. Liu, S. Liu, S. Zhang, B. Li, W. Hu, and Y.-D. Zhang, "SSAU-Net: A Spectral–Spatial Attention-Based U-Net for Hyperspectral Image Fusion," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–16, 2022.

[33] P. Wang, Y. Su, B. Huang, D. Zhu, W. Liu, A. Nedzved, V. V. Krasnoproshin, and H. Leung, "Low-Rank Tensor Completion Pansharpening Based on Haze Correction," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–20, 2024.

[34] R.-C. Wu, S. Deng, R. Ran, H.-X. Dou, and L.-J. Deng, "INF^3 : Implicit Neural Feature Fusion Function for Multispectral and Hyperspectral Image Fusion," *IEEE Trans. Comput. Imaging*, vol. 10, pp. 1547–1558, 2024.

[35] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A Variational Pan-Sharpening With Local Gradient Constraints," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 10 257–10 266.

[36] 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.

[37] 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.

[38] 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.

[39] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on over-sampled multiresolution analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Jan. 2002.