Adaptive Expert Learning for Hyperspectral and Multispectral Image Fusion

Wangquan He, *Student Member, IEEE*, Yixun Cai, Qi Ren, *Student Member, IEEE*, Abuduwaili Ruze, and Sen Jia, *Senior Member, IEEE*

***Abstract*—Hyperspectral image (HSI) and multispectral image (MSI) fusion aims to generate a high-resolution (HR) HSI by leveraging the high spectral fidelity of HSI and the fine spatial details of MSI. However, most existing methods rely on static fusion strategies that assume global consistency in modality contributions, ignoring the inherent regional variability in real-world remote sensing scenes. To address this limitation, we propose an adaptive expert learning framework (AELF) that dynamically models the modal dominance of different regions and adaptively adjusts fusion strategies accordingly. A core component of AELF is the modality-guided complementary module (MGCM), which establishes bidirectional cross-attention pathways between HSI and MSI. It enables each modality to adaptively discover complementary cues across multiple scales while suppressing irrelevant information, providing enhanced feature representation for subsequent fine-grained fusion. Building upon this, we designed the attribute-aware mixture of fusion experts (AMoFE) module, which decomposes the fused features into spectral, spatial, and edge subspaces. Each component is modeled by a specialized expert network, with a soft routing mechanism dynamically adjusting expert contributions based on contextual cues. Extensive experiments on benchmark datasets and a real-world dataset demonstrate that AELF achieves state-of-the-art performance in terms of spectral fidelity and spatial sharpness. Furthermore, our results confirm that the improved data quality brought by the proposed method effectively enhances the overall performance of downstream tasks. The code will be available at: https://github.com/Hewq77/AELF**

***Index Terms*—Adaptive fusion, cross attention, hyperspectral image (HSI) fusion, mixture of experts, multispectral image (MSI).**

I. Introduction

With the advancement of hyperspectral imaging technology, hyperspectral image (HSI) has shown significant potential in land-cover identification and environmental monitoring due to its rich spectral resolution. However, constrained by sensor hardware limitations, the HSI exhibits low spatial resolution, which hinders its performance in high-precision remote sensing tasks. In contrast, the multispectral image (MSI) offers higher spatial resolution but lacks spectral dimensionality. Thus, effective fusion of the MSI and HSI to generate a high-resolution (HR) HSI has become an essential research focus for enhancing remote sensing data quality and supporting downstream high-precision analysis [1], [2], [3].

To achieve high-quality HSI and MSI fusion (HMIF), early studies primarily employed model-driven approaches. These methods typically introduce various prior constraints, such as sparse, low-rank, and nonlocal similarity [4], [5], [6]. By leveraging statistical regularities in specific domains, they design objective functions to guide the reconstruction of HR-HSI. Although effective in specific scenes due to their simplicity and strong physical interpretability, such approaches generally lacked the capacity to model nonlinear relationships and complex spatial structures, making them inadequate for handling the heterogeneous land cover in real remote sensing scenes [7], [8], [9], [10].

With the advancement of deep learning, research in HMIF has gradually shifted from model-driven to data-driven methods. Convolutional neural network (CNN)-based fusion networks can autonomously learn spatial–spectral features from large-scale datasets, significantly improving structural details and spectral consistency in fused images. Building upon this, design strategies such as attention mechanisms, multiscale feature pyramids, and residual fusion architectures [11], [12], [13] have been widely adopted to enhance representational capability. Meanwhile, Transformer-based structures [14], [15], [16], [17] have been introduced to HMIF, leveraging their global modeling strengths to address the limitations of CNNs in capturing long-range dependencies. More recently, state-space methods like Mamba [18], [19], [20] have emerged. Their linear recurrence structures and dynamic channel modeling capabilities offer a balance between performance and efficiency, providing new insights for lightweight and effective multimodal fusion. In addition to these mainstream architectures, diffusion models [21], [22], [23], as a novel generative paradigm, have shown potential in detail reconstruction and spectral consistency through their step-by-step denoising mechanism. Furthermore, graph neural networks (GNNs) and generative adversarial networks (GANs) have also been applied to HMIF tasks [24], [25], [26], offering

Received 22 September 2025; accepted 9 October 2025. Date of publication 13 October 2025; date of current version 27 October 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62271327; in part by the Project of the Department of Education of Guangdong Province under Grant 2023KCXTD029; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011290; in part by Shenzhen Science and Technology Program under Grant RCJC20221008092731042, Grant JCYJ20220818100206015, Grant KQTD20200909113951005, and Grant JCYJ20240813142308012; and in part by the Research Team Cultivation Program of Shenzhen University under Grant 2023JCT002. *(Corresponding author: Sen Jia.)*

The authors are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: hewangquan2022@email.szu.edu.cn; caiyixun2023@email.szu.edu.cn; renqi2023@email.szu.edu.cn; 2350273009@email.szu.edu.cn; senjia@szu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3620897

advantages in modeling spatial structural relationships and enhancing perceptual quality, respectively.

Although fusion performance continues to improve, most existing methods still face critical bottlenecks in fusion strategy design. In HMIF tasks, the two modalities are naturally complementary in spectral and spatial resolution. Traditional fusion approaches generally assume that HSI and MSI are balanced in all regions, and thus adopt a unified and static fusion strategy. However, in real-world remote sensing scenes, this modality complementarity exhibits clear regional heterogeneity rather than global consistency: 1) in areas with simple textures (such as vegetation and water bodies), spectral features are more discriminative, while spatial structures are relatively simple, and spectral information plays a dominant role at this time; 2) in areas with complex textures (such as urban buildings or road edges), spatial textures become the primary discriminative factor due to structural complexity; and 3) at object boundaries, both spectral and spatial changes are abrupt, leading to blurred boundaries and easy aliasing, which makes the fusion results highly sensitive to edge information. Without effectively identifying and modeling such regional differences, fusion outputs are prone to spectral distortion, texture blurring, and edge artifacts, ultimately limiting the generalization ability and practical applicability of the model.

To address this, we propose an adaptive expert learning framework (AELF) for modeling regional modality difference in HMIF. Specifically, the framework adopts a two-stage learning strategy. In the first stage, a lightweight super-resolution (SR) network is used to restore the initial spatial structure of the low-resolution (LR) HSI. Simultaneously, a modality-guided complementary module (MGCM) is introduced to construct bidirectional cross-attention pathways between HSI and MSI at multiple scales. This achieves complementary information enhancement between modalities. Each modality can selectively extract more expressive features from the other, while suppressing irrelevant information, thereby generating high-quality priors for the subsequent fine-grained fusion. In the second stage, we introduce an attribute-aware fusion module [called attribute-aware mixture of fusion experts (AMoFE)] inspired by the mixture-of-experts (MoE) paradigm, which decouples features to be fused into three attribute subspaces: spectral, spatial, and edge. Each subspace is modeled by a customized expert network. A soft routing mechanism dynamically allocates expert responses based on the contextual features of each input region, enabling attribute-level adaptive control. This module is embedded at multiple scales within the fusion encoder, ultimately producing a high-quality fused image with clear structural boundaries. In general, this article makes the following contributions.

1) We propose an AELF that adaptively models the varying modality dominance across different regions in remote sensing scenes. The AELF breaks the conventional assumption of global static consistency. Unlike most existing methods that rely on fixed structures or static fusion strategies, the AELF can dynamically adjust the fusion pathway based on regional content, providing better adaptability to heterogeneous land-cover distributions.

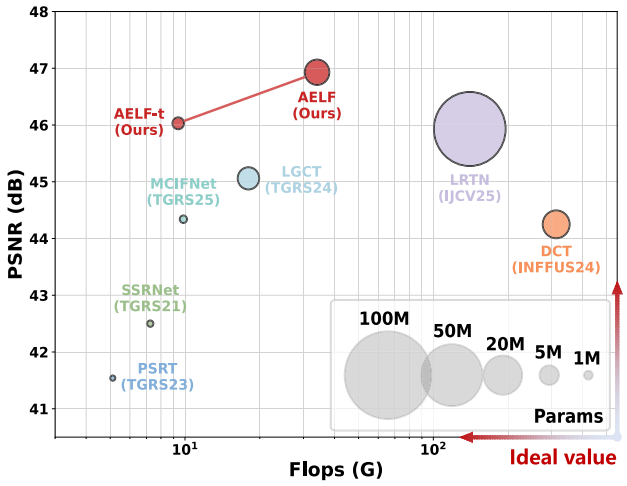

Fig. 1. PSNR–FLOPs comparison of different methods on the Chikusei dataset. *X*-axis: FLOPs; *Y*-axis: PSNR; circle size indicates number of parameters. AELF-t denotes the tiny version of AELF, which achieves excellent performance while significantly reducing the number of parameters.

2) Unlike existing approaches that only consider unidirectional interactions between modalities, we propose an MGCM that leverages a multiscale cross-attention mechanism to establish bidirectional guidance between HSI and MSI features. This interaction allows each modality to adaptively extract discriminative cues from the other, while suppressing redundant or irrelevant information.
3) To enable fine-grained control over the fusion process across diverse scenes, we propose an AMoFE that explicitly decomposes the features to be fused into spectral, spatial, and edge subspaces, each modeled by a dedicated expert. By introducing a content-adaptive soft routing mechanism, AMoFE dynamically adjusts expert responses to different scenes, thereby achieving more flexible and region-sensitive attribute control.
4) Extensive experiments on diverse remote sensing scenes demonstrate that AELF achieves an optimal balance among spectral fidelity, spatial structure preservation, and model efficiency (see Fig. 1). Its tiny version, AELF-t further maintains this balance while significantly reducing the number of parameters, thereby enhancing the practical potential of the proposed method.

The remainder of this article is organized as follows. Section II reviews related work, including model-driven algorithms, data-driven approaches, and recent developments of MoE. Section III presents the overall architecture of the proposed AELF. Section IV reports experimental results and ablation analysis. Section V concludes this article and outlines potential directions for future work.

II. Related Works

A. Model Driven for HMIF

Model-driven approaches primarily rely on prior assumptions about image characteristics. By constructing interpretable optimization models, they achieve fusion reconstruction with a certain degree of physical interpretability and stability.

For instance, Yokoya et al. [27] proposed a coupled matrix factorization method. It separately estimates endmember and abundance information from HSI and MSI, and then fuses the two types of information for reconstruction. Later, researchers incorporated sparse representations and regularization terms [28] to improve model expressiveness. Wu et al. [29] exploited the low-rank structure of HSI in both spectral and spatial domains. They proposed a strategy that combines matrix regression with low-rank reconstruction to enhance fusion quality. Building on these advances, further efforts have been made toward joint spectral–spatial modeling. Dian et al. [30] introduced a fusion method based on spatial–spectral sparse representation, using a joint dictionary to represent features in both domains. Some studies [31], [32] exploit subspace models by incorporating low-rank or deep unfolding priors to enhance HMIF, highlighting the effectiveness of subspace constraints in HSI fusion tasks. In addition, Xu et al. [33] proposed a nonlocal coupled tensor CP decomposition model, which jointly models the high-order structure and nonlocal dependencies of HSI and MSI. Liu et al. [34] embedded nonnegative matrix factorization of HR-HSI into the autoencoder, enabling pixelwise training and the simultaneous estimation of the point spread function and the spectral response function. Although model-driven methods offer advantages in terms of interpretability, their strong reliance on prior assumptions limits their adaptability and nonlinear modeling capabilities in complex land-cover scenes.

B. Data-Driven for HMIF

Compared with traditional model-driven methods, data-driven approaches can automatically learn complex fusion mapping relationships between multimodal data. They offer stronger nonlinear modeling capabilities and better regional adaptability. Thanks to their local modeling ability and end-to-end training characteristics, CNNs have been widely applied in remote sensing image fusion. Xie et al. [35] combined image priors with network structure to achieve high-fidelity fusion through a deeply interpretable architecture. GuidedNet [36] further introduces a HR guidance mechanism to enhance the restoration of structural details. In addition, Sun et al. [37] proposed a deep network with domain transformation modules, which shows improved robustness in complex scenes. Despite these advantages, CNN-based methods are limited by their local receptive fields, which restrict their capacity to model long-range contextual dependencies. To overcome such limitations, recent works [24], [25], [26] have incorporated generative adversarial frameworks and graph-based modeling strategies. These studies aim to capture more complex image relationships by leveraging the expressive power of GANs and GNNs. In addition, related studies [38] have employed diffusion models to learn the spectral distribution of HSI and embed it as a prior into a maximum a posteriori framework, which also improved fusion performance. Transformer-based models have attracted growing attention due to their strong global modeling capabilities. Hu et al. [39] are the first to apply the Transformer to HMIF, using positional encoding to enhance spatial modeling. Subsequently, Jia et al. [16] and Fang et al. [40] extended this approach by introducing multiscale spatial modeling paths for more refined feature fusion. Liu et al. [41] compressed the modeling space using low-rank representations to improve generalization, while LGCT [42] integrates local and global attention collaborative modeling to balance fusion accuracy and computational efficiency. Building on this, the latest studies have introduced state-space modeling mechanisms into HMIF. Zhu et al. [18] refined region-level fusion via collaborative implicit representation, while Li et al. [19] improved robustness to misregistration with an a-Mamba joint framework.

In summary, data-driven methods significantly enhance feature representation and cross-modal modeling through deep network structures. However, most existing methods still adopt a unified fusion strategy and lack the ability to model the region-specific requirements of different land-cover types. As a result, they often struggle to adapt to the diverse spectral and spatial information needs across heterogeneous regions.

C. Mixture of Experts

MoE achieves flexible, sparse, and task-adaptive results by dividing different subnetworks (experts) into independent representation paths and employs a gating mechanism to determine which experts should process each input [43]. A typical MoE architecture consists of multiple expert networks and a router, which dynamically controls the contribution of each expert based on the input. Originally, this mechanism was introduced in language modeling [44]. With the growing demand for sparse modeling, MoE has gradually been adopted in the visual field, especially in general visual models to reduce computational cost and improve generalization [45]. Moreover, MoE has demonstrated strong adaptability and interpretability in low-level vision tasks, such as image fusion and image enhancement [46], [47]. In the field of remote sensing, some studies have explored the potential of MoE for modality selection and feature fusion. For instance, related work [48] has incorporated expert structures into image SR tasks to achieve adaptive modeling of heterogeneous features. In [49], expert networks were used to decompose and perceive the frequency domain to improve the ability to express details. However, these methods do not consider incorporating the highly heterogeneous regional modality characteristics in remote sensing scenes into the design of expert networks. Specifically, they lack an explicit attribute-level functional division and fine-grained modeling in both expert architecture and routing strategy. To this end, this study proposes AMoFE composed of a multiattribute ensemble. By combining structurally diverse experts with a content-adaptive soft routing mechanism, AMoFE enables a more robust and region-sensitive feature fusion.

III. Proposed Approach

This section provides a description of the proposed framework. Sections III-A–III-E elaborate on the problem formulation, overall architecture, core modules, and loss function.

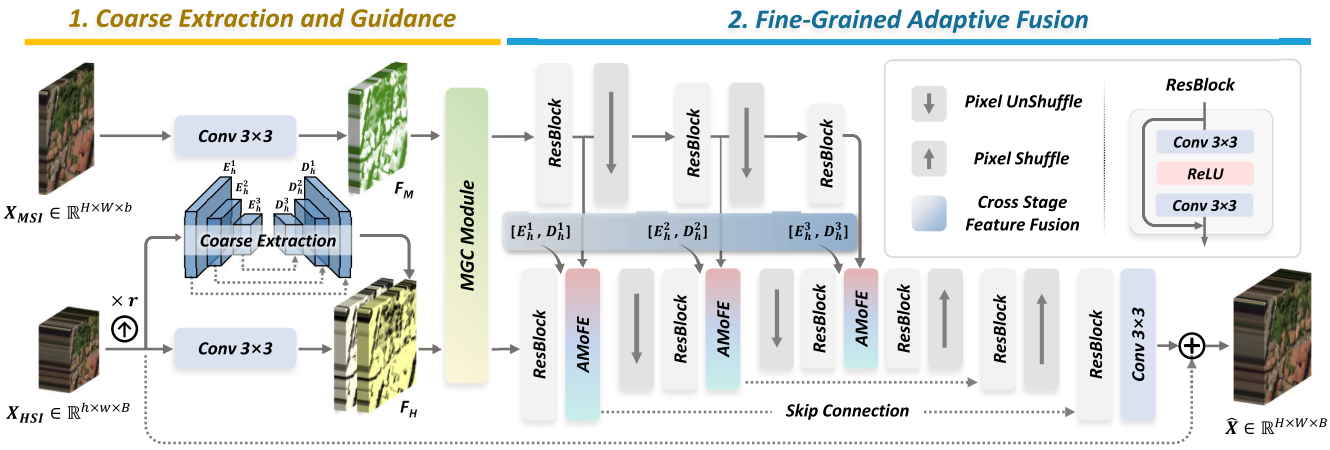

Fig. 2. Overall framework of the proposed AELF. The framework consists of two stages: the first stage is the coarse feature extraction and guidance, which performs preliminary feature extraction from both modalities and introduces the MGCM to enable cross-modal interaction and complementary information modeling. The second stage performs fine-grained adaptive fusion. It embeds AMoFE into a multiscale encoder to enable fine modeling and adaptive integration of spectral, spatial, and edge attributes, ultimately producing high-quality HR-HSI.

A. Problem Formulation

In the task of HMIF, the objective is to reconstruct an HR-HSI from a pair of co-registered observed images: an LR-HSI and an HR-MSI. This inverse problem can be formally expressed as: let $\mathbf{Y} \in \mathbb{R}^{H\times W\times B}$ be the HR-HSI, where $H \times W$ is the spatial size, and B is the number of spectral bands. The observed LR-HSI, defined as $\mathbf{X}_{\text{HSI}} \in \mathbb{R}^{h\times w\times B}$, is obtained by spatial degradation

$$\mathbf{X}_{\text{HSI}} = \mathbf{Y} \downarrow_r \tag{1}$$

where $\downarrow_r$ represents the spatial downsampling with a scaling factor r ($r = H/h = W/w$ and $h \ll H$, $w \ll W$). The other input HR-MSI is defined as $\mathbf{X}_{\text{MSI}} \in \mathbb{R}^{H\times W\times b}$ (where $b \ll B$), which is obtained by spectral degradation

$$\mathbf{X}_{\text{MSI}} = \mathbf{Y} * \mathbf{R} \tag{2}$$

where $\mathbf{R}$ is the spectral response function. This fusion problem is inherently an ill-posed inverse task that simultaneously addresses spatial SR and spectral reconstruction. The core challenge lies in jointly recovering the complete spectral–spatial information of $\mathbf{Y}$ from $\mathbf{X}_{\text{HSI}}$ and $\mathbf{X}_{\text{MSI}}$.

B. Overview

The overall architecture of the proposed method is shown in Fig. 2. It consists of two main stages: a coarse extraction and guidance stage and a fine-grained adaptive fusion stage. These stages are designed to handle low-level structural restoration and high-level attribute representation, respectively, forming a hierarchical and progressive fusion mechanism.

1) Coarse Extraction and Guidance: Initially, we first apply bicubic interpolation upsampling on LR-HSI to obtain the upsampled image $\mathbf{X}^U_{\text{HSI}} \in \mathbb{R}^{H\times W\times B}$, which is fed into the network along with the $\mathbf{X}_{\text{MSI}}$. Subsequently, each input is processed by a 3×3 convolutional layer to extract shallow embeddings. Considering that the interpolated HSI still lacks adequate spatial details, which limits the subsequent fine fusion, we introduce a spatial SR processing process on the $\mathbf{X}^U_{\text{HSI}}$ [50]. The process is implemented using a lightweight U-Net consisting of three encoder stages and three decoder stages. As shown in Fig. 2, each stage is built with a residual block (ResBlock) composed of two 3×3 convolutional layers and one ReLU activation. The encoder produces outputs denoted as $E^n_h (n = 1, 2, 3)$, with step-by-step downsampling achieved via pixel unshuffle. Correspondingly, the decoder outputs are denoted as $D^n_h (n = 3, 2, 1)$, where pixel shuffle is used to restore spatial resolution. Skip connections are introduced between encoder and decoder features of the same scale to prevent information loss and maintain spatial consistency. Supervisory information is further introduced to optimize the training process (see Section III-E for details), thereby enhancing the structural representation of the initial features. Then, the enhanced HSI features are concatenated with the HSI shallow embedded features to form the coarse HSI feature $\mathbf{F}_H$. This process can be expressed as follows:

$$\mathbf{F}_H = \left(\text{Conv}\left(\mathbf{X}^U_{\text{HSI}}\right) \parallel \text{SR}_{\text{HSI}}\left(\mathbf{X}^U_{\text{HSI}}\right)\right) \tag{3}$$

where $\text{SR}_{\text{HSI}}(\cdot)$ represents the SR network and $\parallel$ indicates the channelwise concatenation. To further enhance modality complementarity and cross-modal interaction, we introduce the MGCM. This module takes the coarse HSI fused features $\mathbf{F}_H$ and MSI features $\mathbf{F}_M$ as input and applies a bidirectional cross-attention mechanism across multiple scales to generate enhanced representations from the other modality. The outputs of this interaction are denoted as $\widetilde{\mathbf{F}}_H$ and $\widetilde{\mathbf{F}}_M$, representing the updated features after mutual guidance

$$\widetilde{\mathbf{F}}_H, \widetilde{\mathbf{F}}_M = \text{MGCM}\left(\mathbf{F}_H, \mathbf{F}_M\right). \tag{4}$$

These updated features contain stronger complementary regional information and serve as inputs for the next fusion stage. Through this stage, the network not only recovers low-level structural details but also identifies vital cross-modal complementary regions, providing high-quality priors for attribute decoupling and adaptive fusion in the following stage.

2) Fine-Grained Adaptive Fusion: This stage takes the mutually guided feature pairs from the previous stage as

inputs to perform deep modeling and adaptive fusion. The framework adopts a dual-branch design: the MSI branch is built as a standard encoder, while the HSI branch follows an encoder–decoder structure, with ResBlock as the basic building units. Each ResBlock consists of two 3×3 convolutional layers and an ReLU layer. At each scale n of the HSI branch, the input feature $\mathbf{Z}_H^n$ is processed by an ResBlock. Afterward, to enhance cross-stage feature coherence, we incorporate prior features from the encoder (E_h^n) and decoder (D_h^n) at the same scale in the first-stage SR network. These features are fused after the residual operation, enabling enriched representations

$$\hat{\mathbf{F}}_H^n = \mathrm{Res}\left(\mathbf{Z}_H^n\right) + E_h^n + D_h^n \tag{5}$$

where $n = 1$ and $\mathbf{Z}_H^1$ equals $\widetilde{\mathbf{F}}_H$. Next, the enhanced HSI feature $\hat{\mathbf{F}}_H^n$ and the MSI encoder feature $\widetilde{\mathbf{F}}_M^n$ at the same scale are jointly fed into the AMoFE module to perform attribute-aware dynamic fusion. To build a multiscale fusion architecture, we apply a pixel unshuffle operation to the fused features for downsampling. The resulting output is then used as the input for the next layer

$$\mathbf{Z}_H^{n+1} = \mathrm{US}\left(\mathrm{AMoFE}\left(\hat{\mathbf{F}}_H^n, \widetilde{\mathbf{F}}_M^n\right)\right) \tag{6}$$

where US$(\cdot)$ denotes the pixel unshuffle operation. After being processed by AMoFE modules at each encoder scale, the fused features are transmitted to the decoder via skip connections to enable multiscale information reconstruction. Finally, the decoder output is passed through a convolutional layer and fused with the upsampled LR-HSI, resulting in the final fused image $\hat{\mathbf{X}}$ that matches the spatial resolution of the reference HR-HSI.

C. Modal Complementary Guidance Module

In remote sensing scenes, different regions rely on spectral or spatial information to varying degrees for accurate representation. Therefore, we propose MGCM to achieve feature modeling in regional heterogeneity. Unlike existing works [51], [52] that adopt unidirectional cross attention for intermodality guidance, MGCM establishes a bidirectional guidance pathway between HSI and MSI at multiple scales. This design allows each modality to dynamically extract complementary information from the other prior before fusion, while retaining its own representative advantages. Specifically, the HSI branch focuses on capturing rich textures and edge details from the MSI, whereas the MSI branch is guided by the HSI to enhance spectral discriminability. The details are shown in Fig. 3.

Given HSI and MSI feature maps $(\mathbf{F}_H, \mathbf{F}_M) \in \mathbb{R}^{H\times W\times C}$, the MGCM first applies average pooling at three different scales ($\downarrow 8$, $\downarrow 4$, and $\downarrow 2$) to downsample the input features. This produces multiresolution spatial context information. In each scale branch n, we apply a mutual cross attention (MCA) between the two modalities to enhance their complementary representations. In the MCA module, the cross-modal features are first fed into layer normalization and pointwise convolution. Then, a 3×3 depthwise convolution is applied to capture spatial dependencies across channels. After processing, the features are flattened to $\mathbb{R}^{HW\times C}$. For each scale n, we

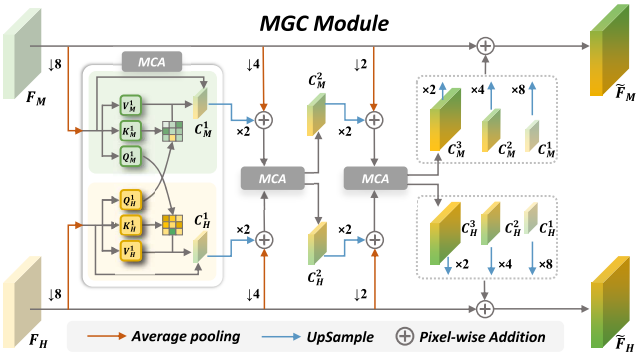

Fig. 3. Architecture of the MGCM. Given coarse features from HSI and MSI, the module employs cross-attention mechanisms across multiple scales to construct intermodal guidance pathways, enabling effective complementary feature modeling.

generate the query (Q), key (K), and value (V) representations. Specifically, from $\mathbf{F}_H^n$, we obtain $(\mathbf{Q}_H^n, \mathbf{K}_H^n, \mathbf{V}_H^n)$, and from $\mathbf{F}_M^n$, we obtain $(\mathbf{Q}_M^n, \mathbf{K}_M^n, \mathbf{V}_M^n)$. Then, the query is exchanged with the key/value to construct the mutually guided attention path

$$\begin{aligned}\mathbf{C}_H^n &= F_H^n + \mathrm{Attn}\left(\mathbf{Q}_M^n, \mathbf{K}_H^n, \mathbf{V}_H^n\right)\\ \mathbf{C}_M^n &= F_M^n + \mathrm{Attn}\left(\mathbf{Q}_H^n, \mathbf{K}_M^n, \mathbf{V}_M^n\right)\end{aligned} \tag{7}$$

where F_H^n and F_M^n represent the outputs of $\mathbf{F}_H$ and $\mathbf{F}_M$ after average pooling at a different scale n. The pair ($\mathbf{C}_H^n$ and $\mathbf{C}_M^n$) is the output of the nth MCA. To achieve guided transfer of cross-scale features, starting from the second scale ($\downarrow 4$), the input of each MCA is fused with the output of the previous layer MCA output after upsampling ($\times 2$). Subsequently, the complementary features at different scales n are upsampled back to the original resolution, aggregated with the backbone input features through skip connections, and finally fused through convolution

$$\widetilde{\mathbf{F}}_{H/M} = \mathrm{Conv}\left(\mathbf{F}_{H/M} + \sum_n \mathrm{US}_n\left(\mathbf{C}_{H/M}^n\right)\right) \tag{8}$$

where US$_n(\cdot)$ denotes the upsampling operation on the nth scale branch. As a result, we obtain the multiscale mutually guided features $\widetilde{\mathbf{F}}_H$ and $\widetilde{\mathbf{F}}_M$, which serve as the inputs for the subsequent fine-grained adaptive fusion stage.

D. Attribute-Aware Mixture of Fusion Experts

In HMIF, different regions exhibit varying degrees of dependence on spectral and spatial information. In addition, edge features are vital for capturing object boundaries and structural transitions, helping preserve spatial clarity and structural consistency in the fused image. However, traditional methods often implicitly embed edge information within spatial modeling, overlooking its independent value in the fusion process. To better accommodate regional differences in reliance on three attributes, we propose AMoFE, which decouples the fused features into three attribute subspaces: spectral, spatial, and edge. Each is modeled by a dedicated expert network. By introducing a soft routing mechanism, AMoFE enables context-aware adaptive expert collaboration, allowing the

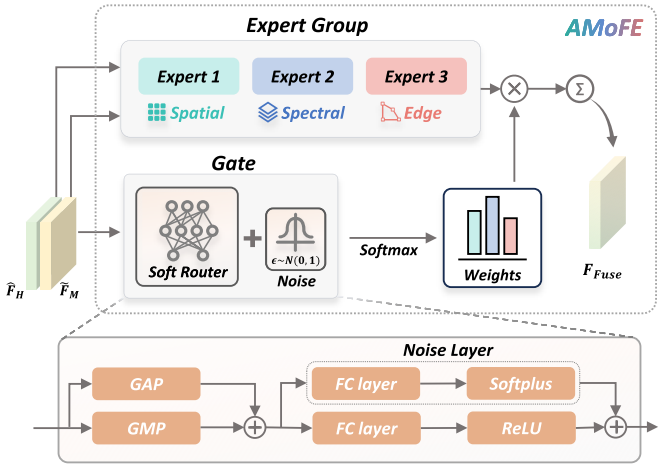

Fig. 4. Architecture of the AMoFE. The fused features from HSI and MSI branches are jointly processed through three attribute-specific experts: spectral, spatial, and edge. A soft routing mechanism then adaptively assigns weights to each expert based on contextual information, enabling region-aware and adaptive fusion.

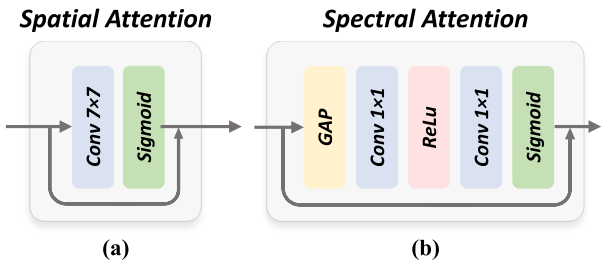

Fig. 5. Architecture of attention modules in AMoFE. (a) Spatial attention module. (b) Spectral attention module.

fusion process to dynamically adjust to the attribute pReferences of different regions. The details are shown in Fig. 4.

1) Expert Group: Given the encoded features from the HSI and MSI branches at scale n be denoted as $(\hat{\mathbf{F}}_H^n, \widetilde{\mathbf{F}}_M^n) \in \mathbb{R}^{H\times W\times C}$, respectively. The processing for each expert is as follows.

1) *Spatial Expert:* The features $\hat{\mathbf{F}}_H^n$ and $\widetilde{\mathbf{F}}_M^n$ are concatenated and fused via convolution operation. Subsequently, a spatial attention module is applied, as illustrated in Fig. 5(a). This module consists of a 7×7 convolution layer followed by a sigmoid activation function. It generates a spatial attention map to reweight and enhance the fused features. Finally, the weighted feature is passed through a convolution to generate the spatial feature $\mathbf{F}_{\text{spa}}^n$

$$\begin{aligned}\mathbf{A}_{\text{spa}}^n &= \text{Conv}\left(\widetilde{\mathbf{F}}_M^n \,\|\, \hat{\mathbf{F}}_H^n\right)\\ \mathbf{F}_{\text{spa}}^n &= \text{Conv}\left(\mathbf{A}_{\text{spa}}^n \cdot \text{SA}_{\text{spa}}\left(\mathbf{A}_{\text{spa}}^n\right)\right)\end{aligned} \tag{9}$$

where $\text{SA}_{\text{spa}}(\cdot)$ denotes the spatial attention module.

2) *Spectral Expert:* Both $\hat{\mathbf{F}}_H^n$ and $\widetilde{\mathbf{F}}_M^n$ are first processed using convolutions and a spectral attention module [53] is then applied to $\hat{\mathbf{F}}_H^n$, as illustrated in Fig. 5(b). The module consists of global average pooling (GAP), two convolutional layers, and an ReLU activation, followed by a sigmoid function to generate the spectral attention map. The attention-enhanced feature is then concatenated with $\widetilde{\mathbf{F}}_M^n$, and the result is further processed by a convolution layer to yield the spectral feature $\mathbf{F}_{\text{spe}}^n$

$$\mathbf{F}_{\text{spe}}^n = \text{Conv}\left(\widetilde{\mathbf{F}}_M^n \,\|\, \left(\hat{\mathbf{F}}_H^n \cdot \text{SA}_{\text{spe}}\left(\hat{\mathbf{F}}_H^n\right)\right)\right) \tag{10}$$

where $\text{SA}_{\text{spe}}(\cdot)$ denotes the spectral attention module.

3) *Edge Expert:* A Laplacian filter is embedded into a convolutional kernel, referred to as Conv_{Lap}, to extract edge responses from $\hat{\mathbf{F}}_H^n$ and $\widetilde{\mathbf{F}}_M^n$. The resulting edge maps are concatenated and passed through a convolution layer to generate the final edge-aware feature $\mathbf{F}_{\text{edge}}^n$

$$\mathbf{F}_{\text{edge}}^n = \text{Conv}\left(\text{Conv}_{\text{Lap}}\left(\hat{\mathbf{F}}_H^n\right) \,\|\, \text{Conv}_{\text{Lap}}\left(\widetilde{\mathbf{F}}_M^n\right)\right). \tag{11}$$

2) Gate Network and Soft Routing: To enable adaptive expert weighting based on regional characteristics, a lightweight gating network is designed to predict soft fusion weights for the three experts. First, GAP and global max pooling (GMP) are applied to the concatenated features

$$\mathbf{z}_{\text{gate}}^n = \text{GAP}\left(\hat{\mathbf{F}}_H^n \| \widetilde{\mathbf{F}}_M^n\right) + \text{GMP}\left(\hat{\mathbf{F}}_H^n \| \widetilde{\mathbf{F}}_M^n\right). \tag{12}$$

Leveraging the global contextual representation constructed by GAP and GMP, the gating network can automatically learning the statistical dependencies between regional content and expert weight allocation during training, thereby enabling implicit modality preference learning without explicit annotations. The aggregated vector $\mathbf{z}_{\text{gate}}^n$ is processed through a fully connected layer and ReLU. A learnable noise component $\eta = \text{SoftPlus}(\text{FC}(\mathbf{z}_{\text{gate}}^n))$ is added to increase routing flexibility. The final fusion scores $\mathbf{s}^n$ are then computed as

$$\mathbf{s}^n = \text{ReLU}\left(\text{FC}\left(\mathbf{z}_{\text{gate}}^n\right)\right) + \eta. \tag{13}$$

Softmax is applied to obtain the fusion weights $\mathbf{w}^n = [w_\alpha, w_\beta, w_\gamma]$. The final fused feature $\mathbf{F}_{\text{Fuse}}^n$ is obtained by the weighted sum of expert outputs

$$\mathbf{F}_{\text{Fuse}}^n = w_\alpha \mathbf{F}_{\text{spe}}^n + w_\beta \mathbf{F}_{\text{spa}}^n + w_\gamma \mathbf{F}_{\text{edge}}^n. \tag{14}$$

E. Training Objective

To jointly optimize multistage fusion reconstruction quality and guide the gating network to maintain moderate imbalance and discrimination in expert selection, we design a composite loss function comprising multiple objectives. Specifically, it includes two image reconstruction supervision terms and a regularization term for the gating distribution. The detailed definitions of each part are as follows.

1) Reconstruction Loss: This part constrains both the output of the first-stage SR network and the final fused image. The ℓ_1 distance is employed to measure the pixelwise discrepancy between the generated image and the reference image, and it is defined as follows:

$$\mathcal{L}_{\text{rec}} = \lambda_1 \left\|\hat{\mathbf{X}}_{\text{SR}} - \mathbf{Y}\right\|_1 + \lambda_2 \left\|\hat{\mathbf{X}} - \mathbf{Y}\right\|_1 \tag{15}$$

where $\hat{\mathbf{X}}_{\text{SR}}$ denotes the output of the SR network in the first stage and λ_1 and λ_2 are weighting coefficients for the two loss terms. For simplicity, they are empirically set to 1.

TABLE I

OVERVIEW OF THE DATASETS USED IN THIS STUDY. SPATIAL RESOLUTION IS GIVEN IN METERS. SPECTRAL RANGE IS IN NANOMETERS. DIMENSIONS ARE REPRESENTED AS $H \times W$ (SPATIAL SIZE IN PIXELS) AND B (NUMBER OF SPECTRAL BANDS)

Datasets	Simulated scenes			Real scene	
	DFC2018	Pavia Center	Chikusei	OHS (HSI)	GF7 (MSI)
Spatial Resolution (m)	1	1.3	2.5	10	2.6
Spectral Range (nm)	380-1050	430-860	343-1018	400-1000	450-900
Dimension (H×W×B)	1202×4172×48	1096×715×102	2517×2335×128	456×385×32	1824×1540×4

2) Coefficient of Variation (CV) Loss: To prevent the gating module from collapsing during training (i.e., models tend to favor the same expert) and to encourage a moderately diverse, we introduce a regularization term based on the CV, defined as follows:

$$\mathcal{L}_{\text{cv}} = \sum_{i=1}^{n} \left(\frac{\sigma(\mathbf{g}_i)}{\mu(\mathbf{g}_i) + \epsilon} \right)^2 \tag{16}$$

where $\mathbf{g}_i$ represents the expert gating vector of the ith scale, $\sigma(\cdot)$ and $\mu(\cdot)$ are the standard deviation and mean, respectively, and ϵ is a small constant (e.g., 10^{-6}) to prevent division by zero.

After combining the above supervision term with the regularization term, the final training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{cv}} \cdot \mathcal{L}_{\text{cv}} \tag{17}$$

where λ_{cv} is the weighting coefficient for the gating regularization term, which is set to 0.01.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

1) Datasets: To test our method under diverse conditions, we used four HSI datasets—three benchmarks and one real-world satellite dataset. Each differs in spatial and spectral resolution, providing a wide range of land-cover complexity from rural fields to dense urban areas. The specific parameters of different datasets are shown in Table I.

1) *DFC2018*[1]*:* The HSI image was provided by the IEEE GRSS Data Fusion Contest 2018 (DFC2018), the dataset consists of HR-HSI acquired by an airborne sensor over the University of Houston and its surrounding urban area.
2) *Pavia Center*[2]*:* The HSI image was captured by the ROSIS sensor over Pavia city center, Italy. The dataset includes nine classes and is widely used in HSI tasks for its compact spatial coverage and high interclass variability.
3) *Chikusei*[3]*:* The HSI image was acquired by the Headwall Hyperspec-VNIR sensor over Chikusei, Japan. It covers rural and semi-urban areas with annotations for 19 classes. The dataset presents challenges due to its high spatial variability and fine-grained class distinctions.
4) *Orbita Hyperspectral Satellite (OHS)-GaoFen-7 (GF7):* The full-resolution dataset consists of co-registered HSI and MSI images captured over the Futian Mangrove area in Shenzhen, China. The HSI was acquired by China OHS, and the MSI was captured by China GF7 satellite.

2) Implementation Details: In the simulated data experiments, we follow the widely used Wald's protocol [54]. The original HR-HSI serves as the reference, while the input HSI and MSI are obtained by downsampling it in both the spatial and spectral dimensions. To simulate the LR-HSI, we first apply a Gaussian filter (kernel size 7×7 and std = 2) to blur the reference image. This is followed by downsampling with scaling factors of 4 and 8, depending on the experimental setting. Meanwhile, the corresponding HR-MSI is generated by applying the spectral response function of Chinese GF-2 satellite to the reference HSI. A 128×128 region of the reference image is cropped as the test sample, while the remaining area is used for training, ensuring no overlap between them. During training, same-size patches are randomly cropped from the training area and fed into the model iteratively. As for real data, we adopt the sampling strategy described in [55] for data processing. Both HSI and MSI are downsampled by factors of 4, following the same processing as in the simulation setting. Supervision is still provided by the original HSI. For the fusion task, a 512×512 pixel region is selected for testing, while the remaining nonoverlapping regions are used for training.

The implementation runs in Python 3.10, using PyTorch 1.13 and MATLAB R2017b. All models are trained on an NVIDIA RTX 6000 GPU with 48 GB of RAM. We use the Adam optimizer with a learning rate of $1e^{-4}$ and train for 10 000 iterations.

3) Quality Metrics: To evaluate the quality of the fused HSI generated by different methods, we adopted six widely used metrics, each capturing a different aspect of spectral or spatial fidelity. These include root-mean-square error (RMSE), peak signal-to-noise ratio (PSNR), relative dimensionless global error in synthesis (ERGAS), correlation coefficient (CC), spectral angle mapper (SAM), and the universal image quality index (UIQI). These metrics reflect how well the fusion preserves both the visual structure and the underlying spectral information. Generally speaking, higher PSNR, CC, and UIQI values indicate better reconstruction performance, while lower RMSE, ERGAS, and SAM values point to reduced distortion or spectral shift. These indicators complement each other and reflect the overall quality of the fused image from different dimensions.

In addition, in real scenes where ideal reference images are unavailable, we introduced no-reference evaluation metrics to further assess the fusion performance [56], including quality with no reference (QNR), D_λ, and D_s. Specifically,

[1]http://hyperspectral.ee.uh.edu/QZ23es1aMPH/2018IEEE/phase2.zip
[2]https://ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes
[3]https://naotoyokoya.com/Download.html

TABLE II

QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON THE DFC2018 DATASET (×4/×8). THE VALUES IN PARENTHESES INDICATE THE IDEAL VALUE FOR EACH METRIC. THE BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Ratio	Metrics	Model-driven		Data-driven						Ours	
		Hysure	LTMR	SSRNet	PSRT	DCT	LRTN	MCIFNet	LGCT	AELF-t	**AELF**
×4	RMSE(0)	2.0595	2.3787	1.3135	0.8680	0.3768	0.3749	0.4549	0.3195	0.2772	**0.2475**
	PSNR(+∞)	43.423	45.576	37.161	40.759	48.008	48.050	46.371	49.441	50.673	**51.658**
	ERGAS(0)	1.0455	1.0723	1.5770	1.0026	0.5228	0.5015	0.6806	0.4051	0.3761	**0.3403**
	CC(1)	0.9882	0.9960	0.9756	0.9962	0.9985	0.9989	0.9972	0.9994	0.9994	**0.9995**
	SAM(0)	1.7530	1.8429	2.8933	1.3105	0.8570	0.8546	0.9859	0.7100	0.6218	**0.5741**
	UIQI(1)	0.9729	0.9885	0.9749	0.9962	0.9985	0.9988	0.9970	0.9994	0.9994	**0.9995**
×8	RMSE(0)	3.2697	3.2517	1.3279	1.0163	0.5705	0.4430	0.4813	0.4585	0.4034	**0.3310**
	PSNR(+∞)	41.770	43.561	37.066	39.389	44.404	46.601	45.881	46.302	47.416	**49.134**
	ERGAS(0)	0.8419	0.7184	0.8146	0.5757	0.3593	0.3021	0.3679	0.2824	0.2638	**0.2307**
	CC(1)	0.9792	0.9934	0.9723	0.9952	0.9977	0.9983	0.9965	0.9988	0.9989	**0.9991**
	SAM(0)	3.5053	1.6962	2.9740	1.7203	1.0770	0.9952	1.0518	0.9604	0.8684	**0.7459**
	UIQI(1)	0.9653	0.9893	0.9726	0.9952	0.9976	0.9982	0.9966	0.9988	0.9989	**0.9991**

QNR provides an overall measure of the fusion quality, with higher values indicating better results. D_λ and D_s reflect the consistency of spectral and spatial details in the fusion process, respectively, with lower values indicating less distortion.

4) Compared Methods: To compare and verify the effectiveness of our method, we selected mainstream methods from two directions: model-driven models and data-driven models. Among the model-driven methods, we considered two representatives. One is LTMR [31], which relies on low-rank tensor regularization; the other is Hysure [57], which is a subspace representation method that excels at maintaining spectral consistency. On the data-driven side, we chose models from several typical architectures. These include CNN-based SSRNet [58], and a series of Transformer-based models such as PSRT [59], DCT [60], LRTN [41], and LGCT [42], reflecting the current exploration direction in image fusion. We also considered MCIFNet [18], which integrates Mamba blocks to capture both local and long-range dependencies more efficiently. To further explore the tradeoff between efficiency and accuracy, we introduce a tiny version of our model, referred to as AELF-t, which adopts a reduced number of parameters while preserving primary structural components. These representative methods cover a wide range of current techniques. Details about their network structures can be found in the corresponding references. To ensure fair comparison, we retrained all models using the parameter settings reported in their original papers.

B. Result Analysis on Simulated Data

1) DFC2018 Scene: We evaluated the performance of various mainstream fusion algorithms on the DFC2018 dataset at 4× and 8× ratios. The quantitative results are summarized in Table II. In general, data-driven methods tend to perform better in terms of spatial representation. However, on this dataset, model-driven methods remain competitive and even outperform certain deep learning models on multiple metrics. This highlights the importance of physical priors, especially in complex land-cover scenes. In contrast, the performance of deep fusion methods on DFC2018 shows obvious variation. Lightweight methods such as SSRNet, PSRT, and MCIFNet underperform compared with other data-driven approaches across most metrics. This reflects their sensitivity to the specific data distribution. Our proposed method and its lightweight version achieve stable and superior performance at both ratios. Notably, they achieve significant gains in PSNR and SAM, indicating strong regional adaptability and scale robustness. Fig. 6 shows the 8× fusion results and corresponding residual maps. The first four methods exhibit visible checkerboard artifacts, while ours better recovers details and edge structures, with smaller residuals, indicating superior spectral consistency and spatial restoration.

2) Pavia Center Scene: The results in Pavia Center scene are presented in Table III. This dataset contains rich urban building areas and clear edge structures, posing higher demands on both spatial detail preservation and spectral consistency. In the 4× fusion task, the traditional method LTMR shows competitive performance across multiple metrics compared with data-driven approaches. However, as the ratio increases to 8×, its performance gap with data-driven methods becomes more evident. This indicates the limited generalization of model-driven strategies at higher scale factors. Overall, LRTN performs robustly under both ratios and achieves the second-best results in most metrics. This suggests its effectiveness in local low-rank modeling. Notably, the lightweight version of our method ranks third on several metrics, maintaining high fusion quality while significantly reducing model complexity. This demonstrates its strong computational efficiency and practical potential. Further 8× visual analysis is shown in Fig. 7, our method performs particularly well in typical urban areas, successfully restoring building edges and material textures, while maintaining low responses in the residual heatmaps. These results confirm the effectiveness of AELF in edge structure reconstruction and spectral consistency.

3) Chikusei Scene: The quantitative results for this scene are presented in Table IV. At both 4× and 8× ratios, the proposed method outperforms others across all major metrics, demonstrating superior spectral reconstruction and structural

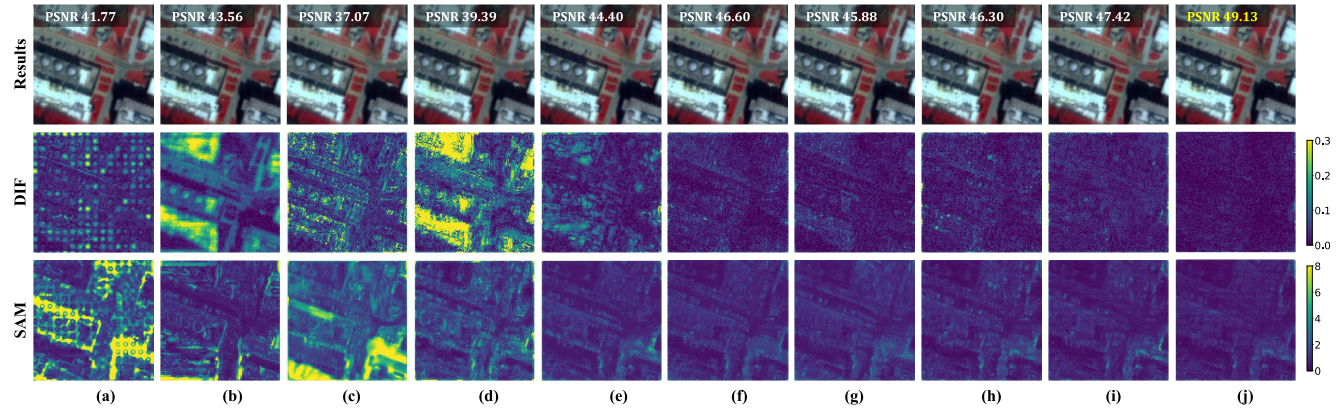

Fig. 6. Visual comparison of different methods on the DFC2018 dataset under ×8 ratio. (Top) Pseudo-color result images generated by each method (RGB from 29/19/9 bands). (Middle) DIF maps showing the mean difference with respect to the reference image. (Bottom) SAM maps representing the spectral angle between the fused results and the reference. (a) Hysure. (b) LTMR. (c) SSRNET. (d) PSRT. (e) DCT. (f) LRTN. (g) MCIFNet. (h) LGCT. (i) AELF-t. (j) AELF.

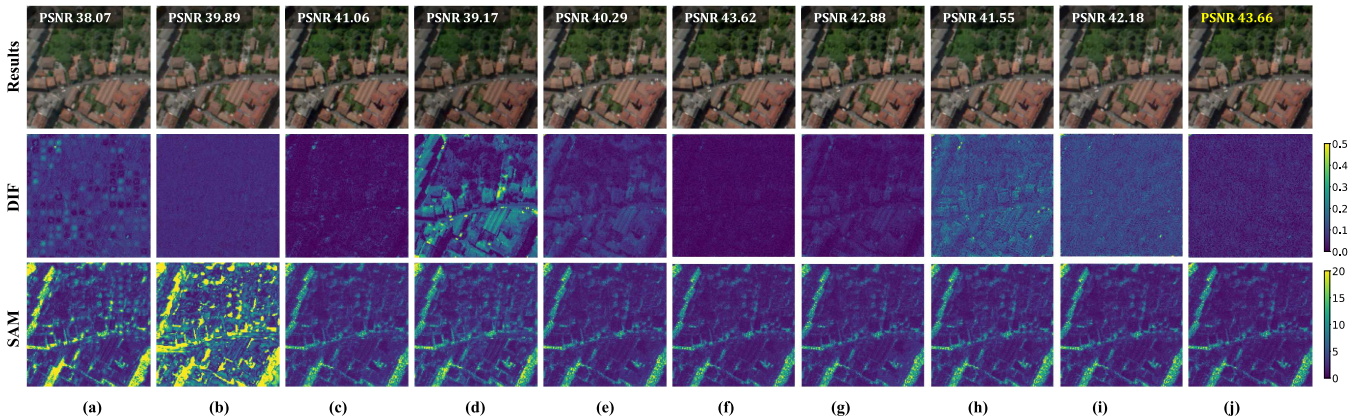

Fig. 7. Visual comparison of different methods on the Pavia Center dataset under ×8 ratio. (Top) Pseudo-color result images generated by each method (RGB from 50/30/15 bands). (Middle) DIF maps showing the mean difference with respect to the reference image. (Bottom) SAM maps representing the spectral angle between the fused results and the reference. (a) Hysure. (b) LTMR. (c) SSRNET. (d) PSRT. (e) DCT. (f) LRTN. (g) MCIFNet. (h) LGCT. (i) AELF-t. (j) AELF.

TABLE III

QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON THE PAVIA CENTER DATASET (×4/×8). THE VALUES IN PARENTHESES INDICATE THE IDEAL VALUE FOR EACH METRIC. THE BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Ratio	Metrics	Model-driven		Data-driven						Ours	
		Hysure	LTMR	SSRNet	PSRT	DCT	LRTN	MCIFNet	LGCT	AELF-t	**AELF**
×4	RMSE(0)	3.1448	3.2772	2.0996	2.5025	1.7534	1.5467	1.7464	1.7416	1.6471	**1.4758**
	PSNR(+∞)	39.171	42.649	41.552	40.028	43.117	44.207	43.152	43.176	43.660	**44.614**
	ERGAS(0)	2.9480	2.8030	2.3260	2.6229	2.0798	1.8969	2.0319	2.0558	1.9654	**1.8285**
	CC(1)	0.9833	0.9816	0.9879	0.9858	0.9900	0.9910	0.9899	0.9899	0.9905	**0.9914**
	SAM(0)	5.0969	5.0704	3.6067	3.9254	3.0289	2.9844	3.2014	3.0577	2.9673	**2.8631**
	UIQI(1)	0.9732	0.9712	0.9872	0.9853	0.9894	0.9904	0.9890	0.9892	0.9899	**0.9908**
×8	RMSE(0)	3.5226	6.5216	2.2212	2.7615	2.4288	1.6549	1.8019	2.1000	1.9527	**1.6471**
	PSNR(+∞)	38.065	39.894	41.063	39.172	40.287	43.620	42.881	41.551	42.182	**43.661**
	ERGAS(0)	1.6236	2.2957	1.2197	1.4224	1.3214	0.9841	1.0350	1.1732	1.1259	**0.9770**
	CC(1)	0.9798	0.9472	0.9869	0.9838	0.9855	**0.9904**	0.9893	0.9875	0.9883	**0.9904**
	SAM(0)	5.4319	9.4169	3.7376	4.2583	3.3844	3.1119	3.2564	3.4844	3.3115	**3.0586**
	UIQI(1)	0.9693	0.9230	0.9860	0.9832	0.9848	0.9897	0.9886	0.9867	0.9876	**0.9898**

consistency. As shown in the qualitative results at 8× (Fig. 8), our method provides sharper reconstructions in typical regions. The SAM residual map also shows deeper contrast, consistent with the quantitative results and highlighting its strength in spectral fidelity. It should be pointed out that model-driven methods exhibit poor scale adaptability on this dataset. While they perform reasonably well at 4×, their performance degrades significantly at 8×. This trend is also evident in

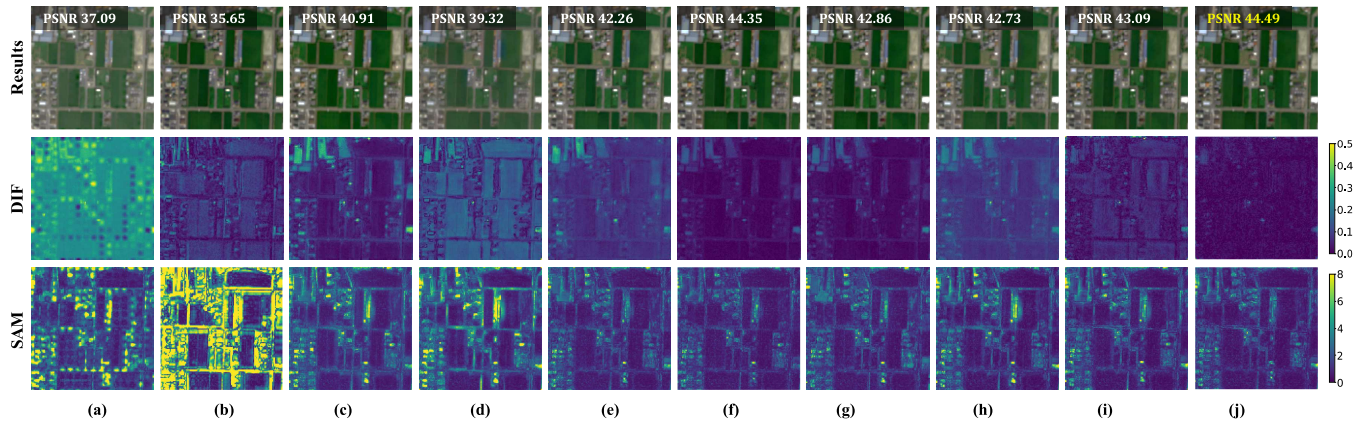

Fig. 8. Visual comparison of different methods on the Chikusei dataset under ×8 ratio. (Top) Pseudo-color result images generated by each method (RGB from 61/41/21 bands). (Middle) DIF maps showing the mean difference with respect to the reference image. (Bottom) SAM maps representing the spectral angle between the fused results and the reference. (a) Hysure. (b) LTMR. (c) SSRNET. (d) PSRT. (e) DCT. (f) LRTN. (g) MCIFNet. (h) LGCT. (i) AELF-t. (j) AELF.

TABLE IV

QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON THE CHIKUSEI DATASET (×4/×8). THE VALUES IN PARENTHESES INDICATE THE IDEAL VALUE FOR EACH METRIC. THE BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Ratio	Metrics	Model-driven		Data-driven						Ours	
		Hysure	LTMR	SSRNet	PSRT	DCT	LRTN	MCIFNet	LGCT	AELF-t	**AELF**
×4	RMSE(0)	3.5800	3.7991	0.7878	0.8795	0.6442	0.5304	0.6372	0.5863	0.5226	**0.4727**
	PSNR(+∞)	38.322	41.669	42.498	41.541	44.246	45.933	44.341	45.064	46.063	**46.934**
	ERGAS(0)	2.7811	3.0144	2.0857	2.0659	1.8192	1.6344	1.8692	1.6593	1.5961	**1.4829**
	CC(1)	0.9864	0.9836	0.9926	0.9930	0.9944	0.9956	0.9937	0.9950	0.9953	**0.9960**
	SAM(0)	2.5790	2.7327	1.6520	1.7335	1.3074	1.1514	1.3140	1.2102	1.0743	**0.9990**
	UIQI(1)	0.9759	0.9740	0.9916	0.9926	0.9938	0.9952	0.9928	0.9947	0.9950	**0.9957**
×8	RMSE(0)	4.0799	8.5038	0.9461	1.1360	0.8100	0.6364	0.7555	0.7674	0.7358	**0.6263**
	PSNR(+∞)	37.085	35.647	40.908	39.319	42.256	44.352	42.862	42.725	43.091	**44.491**
	ERGAS(0)	1.5381	2.5815	1.1940	1.3083	1.0688	0.9130	1.1211	1.0222	0.9867	**0.9127**
	CC(1)	0.9831	0.9391	0.9904	0.9890	0.9918	**0.9939**	0.9908	0.9924	0.9928	0.9937
	SAM(0)	2.9586	5.6513	1.9209	2.2133	1.4745	1.3126	1.4782	1.5408	1.4564	**1.2609**
	UIQI(1)	0.9719	0.9077	0.9899	0.9887	0.9914	0.9934	0.9898	0.9921	0.9925	**0.9935**

the qualitative results, indicating that fixed prior-based models struggle to handle the increased information loss at larger scale factors. Overall, the proposed method has both effective spectral–spatial modeling capabilities and scale robustness, highlighting its potential for application in complex and heterogeneous remote sensing scenes.

C. Per-Band PSNR Analysis

To further evaluate the reconstruction capability across different spectral bands, we present the bandwise PSNR values in Fig. 9. On the DFC2018 scene, the proposed method achieves superior PSNR performance in most bands, demonstrating its strong ability to restore fine details across diverse spectral channels. In contrast, several methods exhibit a noticeable decline in PSNR at the spectral edges, indicating that they suffer from fitting errors due to high-frequency variations. On the Pavia Center and Chikusei datasets, due to the presence of complex urban structures and diverse land-cover types, the PSNR performance differences among methods are more significant across spectral bands. Although the proposed AELF is not consistently superior in every band, it maintains a high level of overall stability. This can be attributed to that the proposed method decouples the fusion features in the spectral, spatial, and edge attributes, which enhances the ability of the model to adapt to region-specific modality dominance and improves its flexibility. In summary, the experimental results demonstrate that the proposed method achieves finer and more reliable fusion representations across varying spectral ranges and object distribution conditions.

D. Result Analysis on Real Scene

To further verify the adaptability and practical application potential of the AELF in real remote sensing scenes, we conducted additional experiments on real datasets, OHS-GF7. Considering that it is impossible to directly obtain the real fusion reference results, we adopted a common degradation verification strategy [41]: degrade and downsample the fusion images generated by different methods to align them with the original LR-HSI, and calculate the SAM error map between them and the reference image to evaluate their spectral consistency. Fig. 10 shows the SAM error map between the degradation map and the reference image of each fusion method. From the results, it can be seen that the proposed method presents a shallower error distribution in most areas,

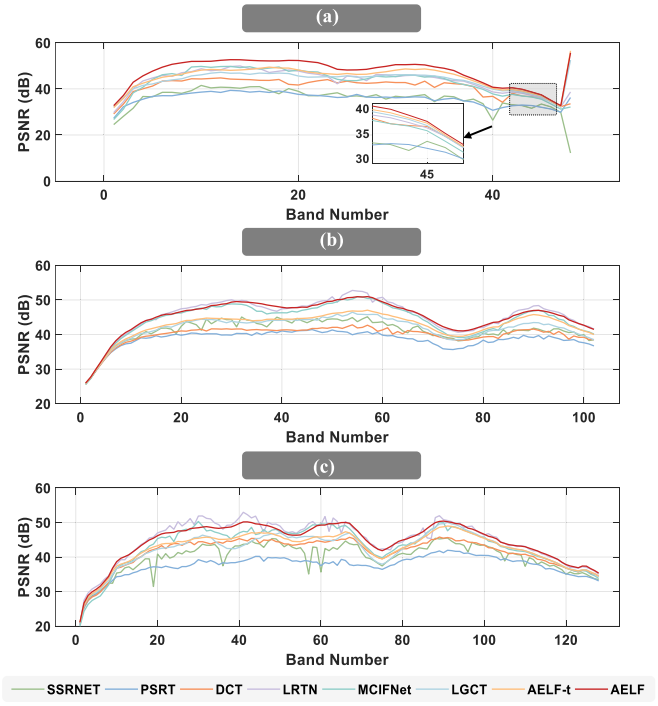

Fig. 9. Per-band PSNR comparison of different methods at ×8 ratio. Higher PSNR values indicate better performance. (a) DFC2018. (b) Pavia Center. (c) Chikusei.

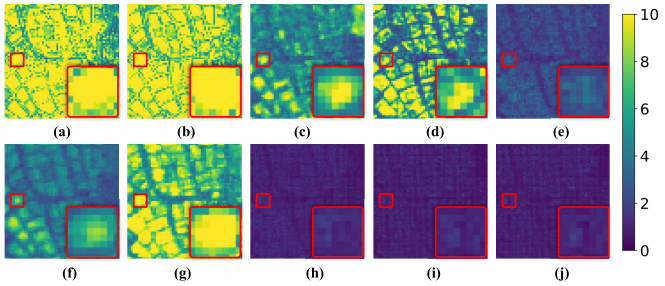

Fig. 10. SAM comparison between the degraded versions of fused results and the reference image on OHS-GF7 real scene. (a) Hysure. (b) LTMR. (c) SSRNET. (d) PSRT. (e) DCT. (f) LRTN. (g) MCIFNet. (h) LGCT. (i) AELF-t. (j) AELF.

reflecting a smaller spectral angle, indicating that it can better preserve the consistency of spectral information between modalities in real scenes and avoid the offset or spectral distortion introduced by fusion. Furthermore, we employed no-reference metrics for quantitative evaluation, and Fig. 11 presents the comparative results of different fusion methods on the OHS-GF7 real scene. It can be observed that AELF achieves superior performance in both QNR and D_λ, indicating its strong capability in preserving spectral information. Although it ranks second in D_s, its overall performance remains significantly advantageous. This trend further demonstrates the practicality and robustness of AELF in real remote sensing scene fusion.

E. Ablation Analysis

1) Effectiveness of Each Modules: To examine the contribution of each core module, we conduct ablation experiments on the DFC2018 dataset under an 8× ratio. The ablation

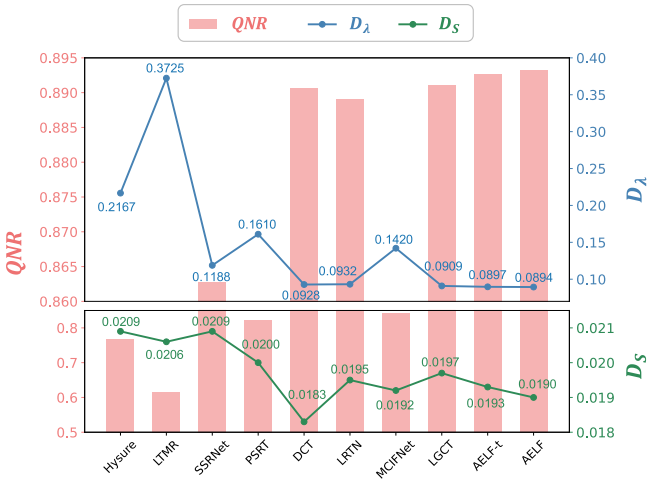

Fig. 11. Comparison of QNR, D_s, and D_λ obtained by different methods on the OHS-GF7 dataset. A higher QNR indicates better performance, while lower D_s and D_λ values represent better results.

TABLE V
PERFORMANCE CONTRIBUTION OF DIFFERENT MODULES IN THE PROPOSED METHOD ON DFC2018 DATASET (×8). BEST RESULTS ARE HIGHLIGHTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Setting	Metrics				
	RMSE↓	PSNR↑	ERGAS↓	SAM↓	UIQI↑
w/o CSFF	<u>0.3405</u>	<u>48.887</u>	<u>0.2322</u>	<u>0.7419</u>	**0.9991**
w/o SR	0.3452	48.769	0.2479	0.7603	0.9989
w/o MGCM	0.3516	48.608	0.2451	0.7823	<u>0.9990</u>
w/o AMoFE	0.3495	48.660	0.2416	0.7714	<u>0.9990</u>
Proposed	**0.3310**	**49.134**	**0.2307**	**0.7459**	**0.9991**

components include the MGCM, AMoFE, the stage-1 SR network, and the cross-stage feature fusion (CSFF). As shown in Table V, the complete model achieves the best performance across all metrics, demonstrating the synergistic effect of all modules in maintaining spectral fidelity and restoring spatial details. Removing CSFF slightly weakens the cross-scale consistency of the model, indicating the usefulness of structural priors in the encoding process of fine fusion. Excluding the stage-1 SR module results in a noticeable degradation in spatial textures, verifying the benefit of coarse-level structure. The significant performance drop occurs when the AMoFE module is removed, especially in SAM and PSNR, emphasizing the importance of attribute decoupling in handling region-specific feature fusion. Meanwhile, the exclusion of MGCM also leads to a certain decrease, confirming that the proposed MGCM facilitates the extraction of complementary information between modalities guidance for fine-grained fusion.

2) Ablation Analysis About MGCM: To verify the effectiveness of the multiscale design in the proposed MGCM, we conduct ablation experiments using different downsampling configurations. As shown in Table VI, the fusion performance improves progressively with the number of scales. Specifically,

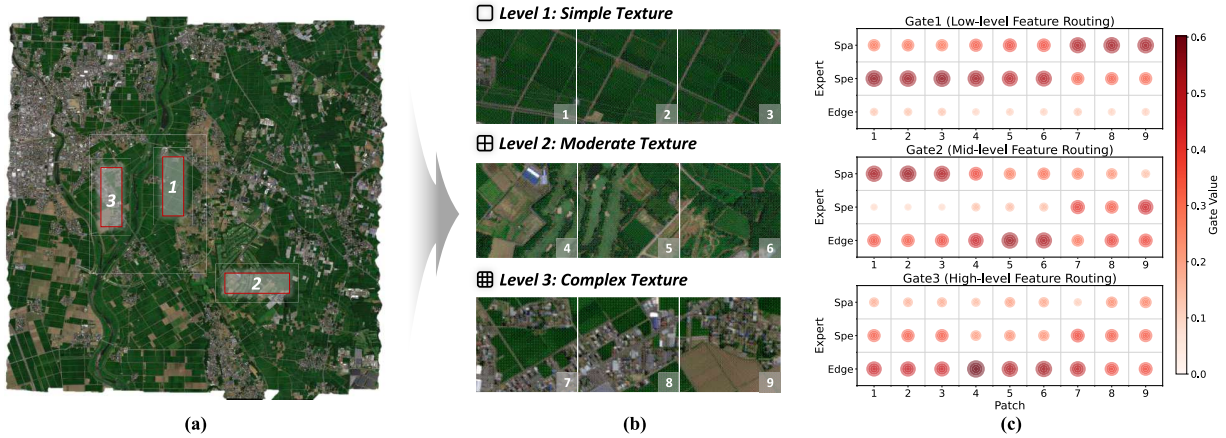

Fig. 12. (a) Chikusei scene with three selected subregions. (b) Patches 1–9 from these subregions, grouped by texture complexity: simple, moderate, and complex. (c) Bubble maps showing the gating weight distributions over the three expert types (spatial, spectral, and edge) for each patch across three gating layers (Gate1–Gate3). Each cell color represents the soft weight of the gating network assigned to an expert at a particular patch.

TABLE VI

ABLATION STUDY WITH CONTRIBUTION OF DIFFERENT DOWNSAMPLING SCALES IN MGCM ON THE DFC2018 DATASET (×8). BEST RESULTS ARE HIGHLIGHTED IN BOLD

Setting	Metrics				
	RMSE↓	PSNR↑	ERGAS↓	SAM↓	#Params
(↓2)	0.3468	48.728	0.2445	0.7586	**6.6529**
(↓2,↓4)	0.3386	48.936	0.2354	0.7587	6.7454
(↓2,↓4,↓8)	**0.3310**	**49.134**	**0.2307**	**0.7459**	6.8378

the single-scale variant (2× only) provides limited cross-modal complementarity and fails to fully capture the semantic diversity across regions. Introduce the medium scale (4×) further enhances spectral fidelity and boundary preservation. With the complete multiscale configuration, the model achieves the best performance across all metrics, indicating that multiscale cross attention effectively enhances complementary features extraction and provides context-aware prior representations that are crucial for the subsequent adaptive fusion process.

3) Analysis About AMoFE: To evaluate the contribution of each attribute-specific expert in the AMoFE module, we conduct ablation experiments based on different combinations of the three experts: spectral (Spe), spatial (Spa), and edge. The results are summarized in Table VII. The full model with all three experts achieves the best performance across all metrics, confirming that the experts play complementary roles in adapting to region-specific feature variations. Furthermore, the two-expert combinations consistently outperform any single-expert variant, indicating that multiattribute modeling enhances the fusion of spectral and spatial information in remote sensing scenes. Notably, the combination of spectral and spatial experts yields the second-best performance, close to the full model, suggesting that these two attributes contribute more dominantly to overall fusion quality, while the edge expert provides useful refinement in structural boundaries.

TABLE VII

ABLATION STUDY WITH DIFFERENT EXPERT COMBINATIONS ON DFC2018 DATASET (×8). BEST RESULTS ARE HIGHLIGHTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Experts			Metrics				
Spa	Spe	Edge	RMSE↓	PSNR↑	ERGAS↓	SAM↓	#Params
✓			0.3690	48.189	0.2491	0.7970	5.8942
	✓		0.3495	48.660	0.2443	0.7825	<u>5.2126</u>
		✓	0.3902	47.703	0.2501	0.8139	**5.8853**
✓	✓		<u>0.3380</u>	<u>48.952</u>	<u>0.2394</u>	<u>0.7505</u>	6.0296
✓		✓	0.3414	48.865	0.2400	0.7572	6.7023
	✓	✓	0.3441	48.796	0.2398	0.7825	6.0207
✓	✓	✓	**0.3310**	**49.134**	**0.2307**	**0.7459**	6.8378

F. Visual Analysis of AMoFE

To further evaluate the ability of the AMoFE to adapt to regional heterogeneity in remote sensing images, we categorize the image into three levels based on local texture complexity: simple texture regions (e.g., grassland), moderate texture regions (e.g., regular farmland grids), and complex texture regions (e.g., dense urban structures). Three representative patches are selected from each level, and their gating distributions are visualized in Fig. 12 to assess the adaptive expert routing of the model under texture variation.

In Gate1, the model mainly activates spectral experts in farmland areas, reflecting a spectral-dominant strategy in early stage feature extraction. In contrast, urban regions show higher activation of spatial experts, indicating an initial preference for spatial features in complex scenes; In Gate2, the preference in farmland shifts toward spatial experts, which is used to extract clearer structural information and field boundaries. Interestingly, urban regions begin to activate spectral experts at this stage, suggesting a mid-level emphasis on material differences. In Gate3, all scene types show stronger activation of edge experts, indicating that the model performs explicit

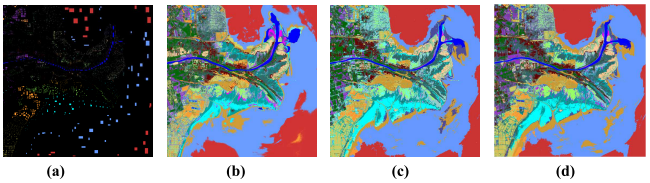

Fig. 13. (a) Ground truth, (b) Ori-MSI (OA = 82.16%), (c) Ori-HSI (OA = 84.48%), and (d) Pre-HSI (OA = 86.27%) result maps on the YRE dataset.

boundary modeling at the final stage. These results demonstrate clear spatial differences and layerwise prioritization in expert selection. The model does not apply a uniform fusion strategy across all regions. Instead, it dynamically adjusts its reliance on different experts according to scene semantics and depth level. For example, in farmland, the model follows a path of "spectral dominance-spatial refinement-edge enhancement," while in urban areas, the pathway becomes "spatial perception-spectral discrimination-edge sharpening." It validates the effectiveness and adaptability of the proposed model in handling remote sensing scene heterogeneity in HMIF.

G. Verification of the Fusion on Classification Task

In this section, we validate the effectiveness of the proposed fusion method in downstream land-cover classification tasks using the Yellow River Estuary (YRE) [61] dataset. The YRE dataset consists of HR-MSI acquired by Sentinel-2 satellite, with a spatial size of 4200 × 4200 and four spectral bands, as well as LR-HSI captured by GF-5 satellite, with a spatial size of 1400 × 1400 and 280 spectral bands. Ground truth covering the eastern part of the Yellow River Delta in China is also provided, involving 20 major land-cover classes, which are spatially consistent with the LR-HSI and reflect the typical complexity of spatial and spectral mixtures.

In the classification experiments, the support vector machine (SVM) was used as a unified classifier to separately perform supervised classification on the original HR-MSI (Ori-MSI), the original LR-HSI (Ori-HSI), and the predicted HR-HSI (Pre-HSI). For each class, 10 pixels were randomly selected as training samples, with the remaining pixels used for testing. Evaluation metrics included overall accuracy (OA), average accuracy (AA), and the Kappa. All classification experiments were repeated 10 times, and the average results are reported. Given that this experiment involves real scene image fusion, we upsampled the ground-truth map by a factor of 3 using nearest-neighbor interpolation to achieve spatial alignment with both the fused image and Ori-MSI [62]. The experimental results (see Table VIII and Fig. 13) show that the Pre-HSI achieves improvements in all evaluation metrics, including OA, AA, and Kappa. The original HSI ranks second, and the MSI performs the worst, which fully validates the effectiveness of the proposed fusion method in improving the accuracy of the downstream land-cover classification task in remote sensing.

H. Model Complexity

To evaluate the tradeoff between performance and computational efficiency of different fusion methods, we draw the PSNR–FLOPs scatter plot as shown in Fig. 1. From the overall distribution, several CNN-based and lightweight Transformer methods achieve competitive performance under low computational cost, reflecting a clear advantage in efficiency. Our method ranks among the top in terms of PSNR, while requiring significantly fewer FLOPs and parameters compared with LRTN and DCT, demonstrating higher computational efficiency. However, compared with other lightweight models such as LGCT and MCIFNet, it remains moderately complex. To further improve efficiency, we propose a tiny variant (AELF-t). It significantly reduces FLOPs and parameter count while maintaining comparable fusion quality, making it more suitable for deployment in resource-constrained environments. The complexity analysis shows that the proposed framework achieves a good balance between performance and efficiency, with the potential to adapt to various application scenes and hardware requirements.

TABLE VIII
CLASSIFICATION ACCURACY RESULTS IN THE YRE DATASET. BEST RESULTS ARE HIGHLIGHTED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Classes	Ori-MSI	Ori-HSI	Pre-HSI
Building	58.14	69.22	64.55
River	99.95	99.91	99.69
Salt Marsh	91.07	67.80	78.91
ShaLLow Sea	89.49	88.27	89.86
Deep Sea	94.82	98.23	98.00
Intertidal Saltwater Marsh	60.11	84.18	72.43
Tidal Flat	65.54	57.24	59.78
Pond	71.59	92.29	86.83
Sorghum	54.20	77.04	79.84
Corn	85.07	78.45	83.32
Lotus Root	59.00	83.76	69.44
AquacuLture	64.29	69.23	70.30
Rice	68.57	62.77	78.16
Tamarix Chinensis	49.37	80.25	92.01
Freshwater Herbaceous Marsh	82.59	92.47	97.15
Suaeda Salsa	94.11	87.96	95.53
Spartina ALterniflora	60.54	89.47	81.48
Reed	56.29	72.65	75.52
Floodplain	78.06	51.34	67.27
Locust	38.05	93.85	84.96
OA (%)	82.16	<u>84.48</u>	**86.27**
AA (%)	71.04	<u>79.82</u>	**81.25**
Kappa	0.795	<u>0.821</u>	**0.842**

V. CONCLUSION

This article proposes a fusion framework named AELF for HMIF, which breaks the assumption of global static consistency in the HMIF problem and enables adaptive perception across different scenes during the fusion process. The method integrates MGCM and AMoFE to perform complementary modeling and dynamic regulation at shallow and deep feature stages, respectively. MGCM enhances the modality complementarity between HSI and MSI through bidirectional cross attention at multiple scales, while AMoFE leverages a soft routing strategy to guide expert networks in collaboratively modeling spectral, spatial, and edge attributes. Experimental results on multiple benchmark datasets show that AELF significantly outperforms existing mainstream methods in both quantitative metrics and visual quality.

In future work, we will focus on lightweight architectural design to reduce computational complexity, improving deployment efficiency on resource-constrained platforms. We also plan to explore task-driven end-to-end fusion strategies to enhance the adaptability of the model to downstream tasks such as segmentation and detection.

REFERENCES

[1] G. Vivone et al., "Deep learning in remote sensing image fusion: Methods, protocols, data, and future perspectives," *IEEE Geosci. Remote Sens. Mag. Replaces Newslett.*, vol. 13, no. 1, pp. 269–310, Mar. 2025.

[2] S. Luo, Y. Qian, L. Bai, Y. Fan, Y. Wang, and W. Kong, "Deep learning-based hyperspectral and multispectral fusion techniques: Review, optimization, and perspectives," *Inf. Fusion*, vol. 124, Dec. 2025, Art. no. 103291.

[3] H.-F. Yan, Y.-Q. Zhao, J. C.-W. Chan, S. G. Kong, N. Ei-Bendary, and M. Reda, "Hyperspectral and multispectral image fusion: When model-driven meet data-driven strategies," *Inf. Fusion*, vol. 116, Apr. 2025, Art. no. 102803.

[4] M. Wang et al., "Tensor decompositions for hyperspectral data processing in remote sensing: A comprehensive review," *IEEE Geosci. Remote Sens. Mag. Replaces Newslett.*, vol. 11, no. 1, pp. 26–72, Mar. 2023.

[5] X. Fu, Y. Guo, M. Xu, and S. Jia, "Hyperspectral image denoising via robust subspace estimation and group sparsity constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5512716.

[6] K. Li, W. Zhang, D. Yu, and X. Tian, "HyperNet: A deep network for hyperspectral, multispectral, and panchromatic image fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 30–44, Jun. 2022.

[7] J. Li, K. Zheng, L. Gao, Z. Han, Z. Li, and J. Chanussot, "Enhanced deep image prior for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, Jan. 2025, Art. no. 5504218.

[8] J. Sun, B. Chen, R. Lu, Z. Cheng, C. Qu, and X. Yuan, "Advancing hyperspectral and multispectral image fusion: An information-aware transformer-based unfolding network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 7407–7421, Apr. 2025.

[9] J. Fang, J. Yang, A. Khader, and L. Xiao, "Deep unfolding network enhanced by transformer priors for unregistered hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Sep. 2024, Art. no. 5531616.

[10] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[11] J. Qu, X. Wu, W. Dong, J. Cui, and Y. Li, "IR&ArF: Toward deep interpretable arbitrary resolution fusion of unregistered hyperspectral and multispectral images," *IEEE Trans. Image Process.*, vol. 34, pp. 1934–1949, 2025.

[12] B. Tu, W. He, Q. Li, Y. Peng, and A. Plaza, "A new context-aware framework for defending against adversarial attacks in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5505114.

[13] C. Zhou, Z. He, J. Dong, Y. Li, J. Ren, and A. Plaza, "Low-rank and sparse representation meet deep unfolding: A new interpretable network for hyperspectral change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, Aug. 2025, Art. no. 5513516.

[14] L. Sun et al., "MDC-FusFormer: Multiscale deep cross-fusion transformer network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Aug. 2024, Art. no. 5528914.

[15] S. Chen, L. Zhang, and L. Zhang, "Cyclic cross-modality interaction for hyperspectral and multispectral image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 1, pp. 741–753, Jan. 2025.

[16] S. Jia, Z. Min, and X. Fu, "Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 96, pp. 117–129, Aug. 2023.

[17] B. Wang et al., "Perceptive spectral transformer unfolding network with multiscale mixed training for arbitrary-scale hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 122, Oct. 2025, Art. no. 103166.

[18] C. Zhu, S. Deng, X. Song, Y. Li, and Q. Wang, "Mamba collaborative implicit neural representation for hyperspectral and multispectral remote sensing image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, Feb. 2025, Art. no. 5504915.

[19] Z. Li, Y. Wen, S. Xiao, J. Qu, N. Li, and W. Dong, "A progressive registration-fusion co-optimization a-Mamba network: Toward deep unregistered hyperspectral and multispectral fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, May 2025, Art. no. 5514815.

[20] Y. Xiao, Q. Yuan, K. Jiang, Y. Chen, Q. Zhang, and C.-W. Lin, "Frequency-assisted mamba for remote sensing image super-resolution," *IEEE Trans. Multimedia*, vol. 27, pp. 1783–1796, 2025.

[21] J. Qu, J. He, Y. Li, W. Dong, and S. Liu, "Progressive synergistic registration and fusion diffusion network for unregistered hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, Apr. 2025, Art. no. 5511614.

[22] Y. Zhong, X. Wu, Z. Cao, H.-X. Dou, and L.-J. Deng, "SSDiff: Spatial–spectral integrated diffusion model for remote sensing pansharpening," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 77962–77986.

[23] J. Zhu, H. Wang, Y. Xu, Z. Wu, and Z. Wei, "Self-learning hyperspectral and multispectral image fusion via adaptive residual guided subspace diffusion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 17862–17871.

[24] Y. Shang, J. Liu, J. Zhang, and Z. Wu, "MFT-GAN: A multi-scale feature-guided transformer network for unsupervised hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, May 2024, Art. no. 5518516.

[25] X. Wang, F. Zhang, K. Zhang, W. Wang, X. Dun, and J. Sun, "Learning spatial–spectral dual adaptive graph embedding for multispectral and hyperspectral image fusion," *Pattern Recognit.*, vol. 151, Jul. 2024, Art. no. 110365.

[26] C. Zhu, S. Deng, Y. Zhou, L.-J. Deng, and Q. Wu, "QIS-GAN: A lightweight adversarial network with quadtree implicit sampling for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5531115.

[27] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[28] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3586–3594.

[29] R. Wu, W.-K. Ma, X. Fu, and Q. Li, "Hyperspectral super-resolution via global–local low-rank matrix estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7125–7140, Oct. 2020.

[30] R. Dian, S. Li, L. Fang, and Q. Wei, "Multispectral and hyperspectral image fusion with spatial–spectral sparse representation," *Inf. Fusion*, vol. 49, pp. 262–270, Sep. 2019.

[31] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5135–5146, Oct. 2019.

[32] J. Liu, D. Shen, Z. Wu, L. Xiao, J. Sun, and H. Yan, "Patch-aware deep hyperspectral and multispectral image fusion by unfolding subspace-based optimization model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1024–1038, 2022.

[33] Y. Xu, Z. Wu, J. Chanussot, P. Comon, and Z. Wei, "Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 348–362, Jan. 2020.

[34] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5522412.

[35] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.

[36] R. Ran, L.-J. Deng, T.-X. Jiang, J.-F. Hu, J. Chanussot, and G. Vivone, "GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution," *IEEE Trans. Cybern.*, vol. 53, no. 7, pp. 4148–4161, Jul. 2023.

[37] W. Sun et al., "Domain transform model driven by deep learning for anti-noise hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Nov. 2023, Art. no. 5500117.

[38] J. Liu, Z. Wu, and L. Xiao, "A spectral diffusion prior for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Aug. 2024, Art. no. 5528613.

[39] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[40] J. Fang, J. Yang, A. Khader, and L. Xiao, "MIMO-SST: Multi-input multi-output spatial–spectral transformer for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5510020.
[41] Y. Liu, R. Dian, and S. Li, "Low-rank transformer for high-resolution hyperspectral computational imaging," *Int. J. Comput. Vis.*, vol. 133, no. 2, pp. 809–824, Feb. 2025.
[42] W. He, X. Fu, N. Li, Q. Ren, and S. Jia, "LGCT: Local–global collaborative transformer for fusion of hyperspectral and multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Nov. 2024, Art. no. 5537114.
[43] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1177–1193, Aug. 2012.
[44] J. Li et al., "CuMo: Scaling multimodal LLM with co-upcycled mixture-of-experts," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 131224–131246.
[45] H. Lin et al., "RS-MoE: A vision—Language model with mixture of experts for remote sensing image captioning and visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, Mar. 2025, Art. no. 5614918.
[46] P. Zhu, Y. Sun, B. Cao, and Q. Hu, "Task-customized mixture of adapters for general image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 7099–7108.
[47] B. Cao, Y. Sun, P. Zhu, and Q. Hu, "Multi-modal gated mixture of local-to-global experts for dynamic image fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 23555–23564.
[48] B. Chen, K. Chen, M. Yang, Z. Zou, and Z. Shi, "Heterogeneous mixture of experts for remote sensing image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–5, 2025.
[49] X. He, K. Yan, R. Li, C. Xie, J. Zhang, and M. Zhou, "Frequency-adaptive pan-sharpening with mixture of experts," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 3, pp. 2121–2129.
[50] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14821–14831.
[51] X. Cao, Y. Lian, K. Wang, C. Ma, and X. Xu, "Unsupervised hybrid network of transformer and CNN for blind hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Jan. 2024, Art. no. 5507615.
[52] X. Cao, Y. Lian, J. Li, K. Wang, and C. Ma, "Unsupervised multi-level spatio-spectral fusion transformer for hyperspectral image super-resolution," *Opt. Laser Technol.*, vol. 176, Sep. 2024, Art. no. 111032.
[53] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.
[54] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.
[55] T. Huang, W. Dong, J. Wu, L. Li, X. Li, and G. Shi, "Deep hyperspectral image fusion network with iterative spatio-spectral regularization," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 201–214, 2022.
[56] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogrammetric Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.
[57] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.
[58] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021.
[59] S.-Q. Deng, L.-J. Deng, X. Wu, R. Ran, D. Hong, and G. Vivone, "PSRT: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503715.
[60] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Reciprocal transformer for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102148.
[61] K. Ren, W. Sun, X. Meng, G. Yang, J. Peng, and J. Huang, "A locally optimized model for hyperspectral and multispectral images fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5519015.
[62] H. Yu, Z. Ling, K. Zheng, L. Gao, J. Li, and J. Chanussot, "Unsupervised hyperspectral and multispectral image fusion with deep spectral–spatial collaborative constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Oct. 2024, Art. no. 5534114.

Wangquan He (Student Member, IEEE) received the M.S. degree in information and communication engineering from Hunan Institute of Science and Technology, Yueyang, China, in 2022. He is currently pursuing the Ph.D. degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image fusion and super-resolution.

Yixun Cai received the B.S. degree from Shenyang Jianzhu University, Shenyang, China, in 2023. He is currently pursuing the M.S. degree in computer science and technology with Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image fusion and super-resolution.

Qi Ren (Student Member, IEEE) received the B.S. degree in computer science and technology from Shanxi Datong University, Datong, China, in 2020, and the M.S. degree in information and communication engineering from Hunan Institute of Technology, Yueyang, China, in 2023. He is currently pursuing the Ph.D. degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image processing and spectral construction.

Abuduwaili Ruze received the B.S. and M.S. degrees in software engineering from Xinjiang University, Ürümqi, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image classification and image segmentation.

Sen Jia (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively.

Since 2008, he has been with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, where he is currently a Full Professor. His research interests include remote sensing image processing, signal and image processing, and machine learning.