MHIAIFormer: Multihead Interacted and Adaptive Integrated Transformer With Spatial-Spectral Attention for Hyperspectral Image Classification

Delong Kong, Jiahua Zhang, Shichao Zhang, Xiang Yu, and Foyez Ahmed Prodhan

***Abstract*—Deep learning is an effective method for hyperspectral image (HSI) classification, where CNN-based and Transformer-based methods have achieved excellent performance. However, there are some drawbacks to the existing CNN-based and Transformer-based HSI classification approaches: 1) CNN-based methods are deficient in showing the extraction of multiscale features and localized features owing to the fixed-size input patch. 2) the MHSA module ignores the interaction capability between multiple attention heads, which leads to insufficient feature fusion in various directions. 3) The weights of attention heads in various directions are disregarded in the MHSA and attention heads are simply concatenated horizontally. To address the above-mentioned limitations, a novel multihead interacted and adaptive integrated transformer (MHIAIFormer) with spatial-spectral attention, which integrates the respective advantages of convolutions and transformers is proposed in this study. A pyramidal spatial-spectral attention (PS2A) feature extraction module is adopted to efficiently capture the localized and multiscale feature information of HSI. The output of PS2A is then sent to the transformer encoder stage through a grouped multiscale cross-dimension embedding module, which includes additive self-attention using multihead interaction and MHSA with adaptive multihead merging to capture the long-range dependencies of the features. Extensive experiments on four datasets verify that our proposed approach achieves more satisfactory classification accuracy when compared with state-of-the-art models. The overall accuracy of the proposed model achieved 95.97%, 98.68%, 92.68%, and 99.49% on four datasets.**

***Index Terms*—Deep learning (DL), hyperspectral image (HSI) classification, multihead interacted and adaptive integrated transformer (MHIAIFormer), multihead self-attention (MHSA).**

NOMENCLATURE

$X \in R^{1\times C\times P\times P}$	Input HSI patch.
$X \in R^{B\times P\times P}$	Feature map of PS2A output.
M_{spa}	Multiscale feature map of spatial branch.
M_{spe}	Multiscale feature map of spectral branch.
F	Output feature map after Conv2D in PS2A.
F_{spa}	Output of the spatial branch.
F_{spe}	Output of the spectral branch.
F_{SAWM}	Output of the SAWM.
F_k	Enter the feature map of the kth MATM.
$\hat{F}_k$	Feature map of the first four branches merged in the kth MATM.
$W_h,\ W_w$	Attention distributions in horizontal and vertical direction.
M_{MHIA}	Output of the MHIA.
X_{out}	Output of the MHIASA.
Z_i	Attention head i in AMHCat.
$\hat{Z}_i$	Output for each head Z_i in AMHCat.
Z	Final output of AMHCat.
$f^{1\times1}(.)$	1×1 convolution layer.
$f(.)$	Linear layer.
$\sigma(.)$	Sigmoid function.
$\delta(.)$	ReLU activation function.
$\odot$	Hadamard product of matrices.

Manuscript received 4 May 2024; revised 2 July 2024; accepted 1 August 2024. Date of publication 9 August 2024; date of current version 26 August 2024. This work was supported in part by the Central Guiding Local Science and Technology Development Fund of Shandong-Yellow River Basin Collaborative Science and Technology Innovation Special Project under Grant YDZX2023019, and in part by the Natural Science Foundation of Shandong Province under Grant 2018GNC110025 and Grant ZR2020QF067. *(Corresponding author: Jiahua Zhang.)*

Delong Kong, Shichao Zhang, and Xiang Yu are with the Remote Sensing Information and Digital Earth Center, College of Computer Science and Technology, Qingdao University, Qingdao 266071, China (e-mail: kongdelong@qdu.edu.cn; 2021010034@qdu.edu.cn; 2020010031@qdu.edu.cn).

Jiahua Zhang is with the Remote Sensing Information and Digital Earth Center, College of Computer Science and Technology, Qingdao University, Qingdao 266071, China, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: zhangjh@radi.ac.cn).

Foyez Ahmed Prodhan is with the Department of Agricultural Extension and Rural Development, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur 1706, Bangladesh (e-mail: foyez@bsmrau.edu.bd).

Data is available on-line at https://github.com/Delon1364/MHIAIFormer.

Digital Object Identifier 10.1109/JSTARS.2024.3441111

I. INTRODUCTION

HYPERSPECTRAL images (HSIs) contain dozens or even hundreds of consecutive bands ranging from visible light to infrared light in each pixel. Applications for HSIs are numerous due to the copious spectral and spatial information, such as mineral survey [1], medical imaging [2], water quality monitoring [3], observation of urban development [4], and fine-grained classification of crops [5]. The most basic and important task is classification, i.e., assigning a specific category to each pixel.

A series of traditional HSI classification methods with hand-extracted features have been developed to concentrate on the spectral information of HSI [6], [7], [8], [9], etc. To consider the relevance of spatial contextual information, many different methods have also been applied [10], [11], [12], [13]. However, these traditional methods are constrained to capture shallow features considering their reliance on human feature extraction

and prior information. In addition, their inadequate generalization capacity results in poor classification and a lack of robustness.

A surge in deep learning (DL)-based HSI classification has been brought about by the growing popularity of DL methods. DL models [14], [15], [16], [17] can automatically extract features and obtain more abundant semantic information than traditional methods. Among these DL methods, CNNs have proven their excellent classification performance by being employed in numerous HSI classification projects. Several CNN-based methods [18], [19], [20], [21], [22], [23] have all yielded favorable results in classifying HSI. Although these CNN-based architectures can achieve positive success in classification results by stacking multiple layers to learn information from a larger receptive field, they still lack global connectivity. Furthermore, CNNs focus too much on spatial information, leading to unavoidable shortcomings in revealing the spatial-spectral dependencies of HSI.

Recently, transformers have shown promising results in various vision tasks, e.g., image classification [24], objection detection [25], and instance segmentation [26]. Given the self-attention mechanism's properties, the transformer can gather global characteristics from vast amounts of data to produce a more robust and precise representation than CNN's inductive bias through long-distance modeling and adaptive spatial aggregation [27]. Transformers have been investigated in several studies for HSI applications [28], [29], [30]. SF [31] generates local embeddings by grouping spectral dimensions in neighboring bands of HSI. GAHT [32] applies the embedding method that applies a group convolution for grouping along the channel dimensions and reduces the model complexity through a hierarchical structure. MCAL [33] employs spectral attention to compress the number of spectra and provides a novel cross-layer fusion technique to represent characteristics at various layers. MSTNet [34] uses a pure transformer encoder and decoder structure to ensure the feature extraction process has a global field of view. DCN-T [35] devises a new tri-spectral image generation method to make the model closer to that of the three-channel image task in the transformer and combines it with pixel similarity-based clustering to extract contextual information. LESSFormer [36] proposes a novel locally enhanced spatial-spectral joint method for HSI classification, which converts HSI into an adaptive spatial-spectral joint token representation and further improves the recognition capability of these tokens by capturing both local and global information. SCFormer [37] introduced an innovative and compact Transformer network to investigate spectral coordinate priors, reducing spectral position disturbance from the convolution process and enhancing algorithm stability. These transformer networks benefit from the extraction of global dependencies to obtain satisfactory classification results. However, these transformer methods are biased towards capturing spectral information at the expense of spatial feature extraction. In addition, the transformer structure is still inferior to CNNs with smaller data volumes due to the lack of proper inductive bias.

The systematic combination of convolutions and transformers has received much attention in building high-performance networks, which can fully utilize the spectral-spatial information of HSI. SSFTT [38] uses 3-D and 2-D convolution to extract spatial-spectral joint features and introduces a Gaussian-weighted feature tokenizer to quickly extract high-level semantic information. BS2T [39] employs a three-phase CNN bottleneck structure to separately extract the spatial and spectral information and examine the relationship between position-spatial, spatial-spectral, and spatial-spatial by adding spatial-position coding and spectral information to multihead self-attention (MHSA). MorphFormer [40] develops a morphological network architecture that combines the MHSA with a morphological spatial-spectral convolution module after initially extracting features using a CNN backbone network. HiT [41] combines 3-D convolution in conjunction with an attention mechanism to first extract spatial-spectral fusion feature representations and then capture subtle spectral differences and spatial contextual information using a transformer.

Although these transformer networks incorporating convolutions have demonstrated impressive performance, there are still some challenges in the current networks. The transformer is prone to overfitting when the number of training samples is limited and performs poorly when it comes to explicitly extracting local and multiscale features. Besides, the transformer exploits the MHSA to acquire global contextual information from all tokens, but it does not take into account how each attention head interacts with the others. Moreover, when each attention head is combined in the MHSA, the weights of the attention heads from various directions are disregarded, and the entire output of all the attention heads is integrated by merely stacking them horizontally.

To address the aforementioned problems, in this study, a multihead interacted and adaptive integrated transformer (MHIAIFormer) with spatial-spectral attention is proposed to be applied to the classification of HSI, which integrates the respective advantages of convolutions and transformers and utilizes a few labeled samples. In MHIAIFormer, we design a pyramidal spatial-spectral attention (PS2A) feature extraction module to effectively retrieve local and multiscale features. In addition, we design a productive grouped multiscale cross-dimension (GMCD) embedding module to efficiently extract spatial context information while preserving abundant spectral information. Multihead interacted additive self-attention (MHIASA) is proposed to effectively compensate for the missing spectral dimension modeling capability and multiattention head interaction capability of MHSA. The adaptive multihead catenation (AMHCat) module, as an improvement to MHSA, can attain the attention weights of features in various directions adaptively. In general, the main contributions of this work are as follows.

1) We propose an MHIAIFormer network for HSI classification, a network that absorbs the respective advantages of convolutions and transformers.
2) A PS2A is designed to retrieve shallow features in both spatial and spectral dimensions effectively, and the spatial adaptive weights module (SAWM) is used to perform the feature fusion of the two dimensions.
3) An efficient embedding module called GMCD is proposed to extract the spatial information in the form of grouping

while preserving the spectral feature through a multiscale asymmetric triplet attention module (MATM).

4) We produce a MHIASA module to compensate the multihead interaction capability of MHSA and improve the integration mechanism of each attention head in MHSA by using AMHCat.
5) Experimental results on four benchmark datasets show that MHIAIFormer outperforms other state-of-the-art DL methods in classification performance.

The rest of this article is organized as follows. Section II gives the related work on the application of attention mechanisms and transformer-based networks for CV. Section III describes the proposed network structure. Section IV describes the dataset description, the experimental setup, and the analysis of the experimental results. Section V provides a summary of this study.

II. Related Works

II. Attention Mechanism

Attention is a mechanism of the human brain to solve information overload by selecting a small portion of useful information to focus on from a large amount of input information. When DL processes vast amounts of input data, it can also leverage the attention mechanism of the human brain to enhance model efficiency. As the most noteworthy core technology in DL, the attention mechanism is widely used in various fields such as natural language processing and computer vision (CV). In the field of CV, the "squeeze-and-excitation" (SE) [42] module and the convolutional attention mechanism module (CBAM) [43] effectively explore the attention of channels and spatial locations to generate attention weight distributions automatically. Coordinate attention [44] increases the representation of interesting objects by embedding position information into channel attention, maintaining precise position information along one spatial direction while capturing long-distance dependencies along the other, and obtaining an attention map that contains both direction-aware and position-sensitive attention. Triplet attention [45] is a low-cost, high-efficiency attention mechanism that generates cross-dimensional dependency distributions by modeling spatial and channel attention in nearly parameter-free ways. Self-attention mechanisms are often practiced in the field of CV as well. To improve the modeling capability, self-attention models often adopt the query-key-value mode, which maps the input $X \in R^{n\times d}$ into three matrices $W_Q,\ W_K,\ W_V$. The process of generating the attention weight map involves computing the similarity between every token vector and applying the softmax function to normalize the outcome. This function can be explained as follows:

$$\text{Attention}\ (Q,\ K, V) = \ \text{softmax}\left(\frac{Q\cdot K^T}{\sqrt{d}}\right)\cdot V. \quad (1)$$

However, as the number of pixels in the image increases, the computational effort of the self-attention module also increases. A few studies have been implemented to enhance the MHSA and reduce the computational burden. ELS2T [46] constructs an efficient lightweight network by applying the lightweight separable spatial-spectral self-attention module instead of MHSA. SwiftFormer [47] removes the key-value interaction from the attention mechanism by investigating the efficient additive self-attention (EASA) module, which explores the interaction between query-key through additive attention learning the relationship between tokens. We improved the EASA to make it more suited for HSI processing. Furthermore, current attention mechanisms that are employed as feature extraction modules tend to focus only on local features and neglect the interaction between global and multiscale features. Accordingly, we construct PS2A in this work to learn abundant multiscale information while concentrating attention.

B. Vision Transformer

In CV, transformers are starting to shine. When Vision Transformer (ViT) [48] was initially suggested as a tool for image classification, it sparked an explosion of ViT innovations. ViT provides a new design paradigm to transform the original input image into a series of blocks, which are then mapped to specific embedding dimensions by a trainable linear layer that feeds into a transformer encoder to model global long-range dependencies. However, ViT ignores local features, and since there is no inductive bias, it requires a large amount of data to avoid overfitting. Various design strategies have been explored to incorporate the benefits of convolutional neural networks into transformer models to improve performance. Mobile-Former [49] is a network architecture for parallel processing of MobileNet and transformer, incorporating the advantages of MobileNet for local processing and transformer for global interaction. CMT [50] embeds deep convolution into the transformer encoder to enhance the local information while obtaining long-range modeling relationships. In addition, some network architectures achieve better performance by adapting the structure of self-attention. DHVT [51] successfully bridges the performance gap between CNNs and ViTs by introducing a new "head token" in MHSA to help recalibrate the channel representations and make the representations of different channel groups interact with each other. SMT [52] blends multiscale convolution and achieves information fusion across different heads, which can effectively capture the transition of features from local to global dependencies, thus yielding superior performance. However, these transformer-based networks still lack statistics on the attention weights from each head or the weights of feature information in various directions. When building ViT models for high-dimensional HSI classification tasks, the advantages of the transformer continue to be problematic since they need to pay attention to the information interaction between each attentional head.

C. Hybrid Models Based on CNN and Transformer

CNNs and Transformers combined in a methodical way have drawn a lot of interest in the construction of high-performance networks that can fully exploit the spectral-spatial information of HSI. Bai et al. [53] initially compressed spectral information using 1-D convolution, then incorporated spatial location information into self-attention to obtain location attention weights

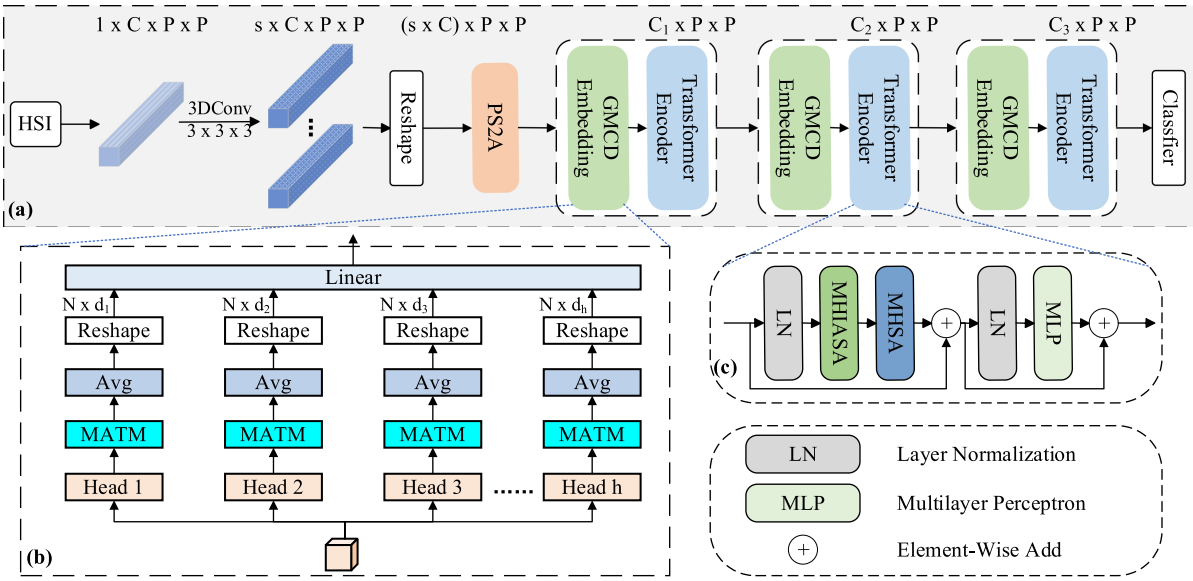

Fig. 1. Architecture of the proposed MHIAIFormer for HSI classification. (a) Overall pipeline. (b) Illustration of the GMCD, which consists of MATM. (c) Architecture of the transformer encoder.

and relabeled the HSI, and applied a mask prediction branch to obtain end-to-end output. HybridFormer [54] utilized CNN to extract shallow features and designed a spectral-spatial attention to replace the original self-attention mechanism in transformer. CTMixer [55] used group bottleneck residual convolution to extract low-level semantic information, which is later fed to transformer encoder with convolution. PyFormer [56] divides the sequence after PCA into hierarchical segments representing different levels and uses a convolutional pyramid-like structure to extract information at different levels, and Transformer is applied independently to each level of the hierarchy. CTN [57] utilized PCA to select the spectra and then added position information before entering the convolutional transformer that combines convolution and self-attention. Arshad and Zhang [58] applied a combination of 3-D and 2-D convolution for feature extraction following PCA. They additionally employed interleaved patterns of local and hierarchical attention blocks in Transformer to efficiently capture short and long-range spatial dependencies and model cross-window interactions. The above-mentioned methods utilize PCA or grouped convolution for low-level feature extraction, which destroys the spectral continuity of HSIs. The transformer branching in the hybrid model does not consider the interaction between attention heads and the effective fusion of different directions. Therefore, we construct the PS2A, MHIASA, and AMHCat modules to address the above-mentioned issues.

III. Methodology

We are committed to developing a deep network combining transformer and convolution that completely considers the spatial and spectral characteristics of HSI and guarantees that it will perform better in terms of classification and accuracy.

For the input HSI patch $X \in R^{1\times C\times P\times P}$, where P stands for height and width, and C represents the number of spectra. To ensure that the extracted feature maps retain both spatial semantic features and spectral dependency information, X is first subjected to a 3-D convolution to initially extract the features and reduce the spectral dimensionality, yielding $X \in R^{s\times C\times P\times P}$. To accommodate the next 2-D convolution, we reshape the size of X to $(s \times C) \times P \times P$, and then proceed to the PS2A module, which exploits the effectiveness of convolutional and attentional mechanisms for localized feature extraction to ensure that abundant tokens and more adequate shallow features are obtained. Then, three stages of transformer-based are entered for mid-level and deep-level feature extraction. In the transformer block, the GMCD and MHIASA are designed to obtain the interactions between multiple heads more efficiently, and we improve the fusion between different heads in MHSA by AMHCat. Finally, a linear is utilized for classification. The overall structure of the model is shown in Fig. 1.

A. PS2A Feature Extraction

1) PS2A Feature Extraction: For the PS2A as shown in Fig. 2, the feature map $X \in R^{B\times P\times P}$, where $B = s \times C$ first passes through a 2-D convolution layer followed by a batch normalization (BN) layer and a rectified linear unit (ReLU) activation function to obtain a preliminary feature map containing spatial and channel information. Then, the output feature map F enters the spatial and spectral branches, respectively.

To extract abundant multiscale spectral-spatial information, PS2A introduces a pyramidal convolution [23] module divided into spatial and channel kernels, which characterize the local relationships in spatial and spectral dimensions, respectively. In addition, the multiscale convolution kernel only fluctuates in one dimension to reduce the model's computational complexity. Nevertheless, the extracted multiscale features lack the important distribution of the features in that dimension, which leads to slow feature extraction information. Therefore, intending to

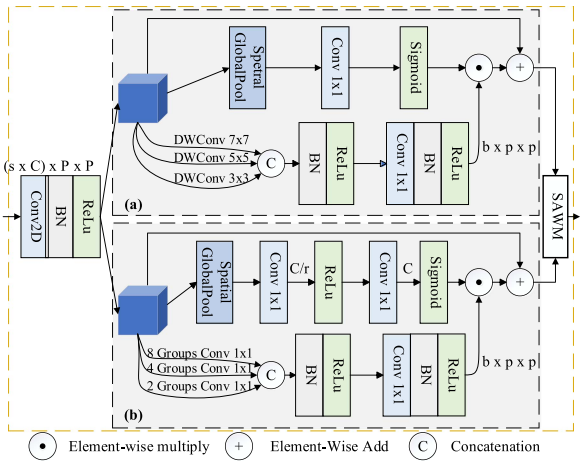

Fig. 2. Architecture of the PS2A. (a) Pyramidal spatial attention branch. (b) Pyramidal spectral attention branch.

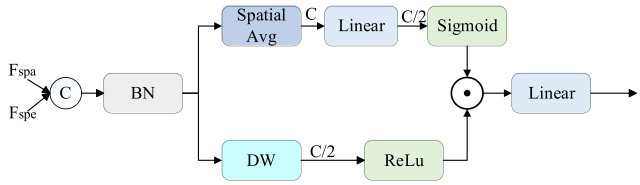

Fig. 3. Illustration of the SAWM.

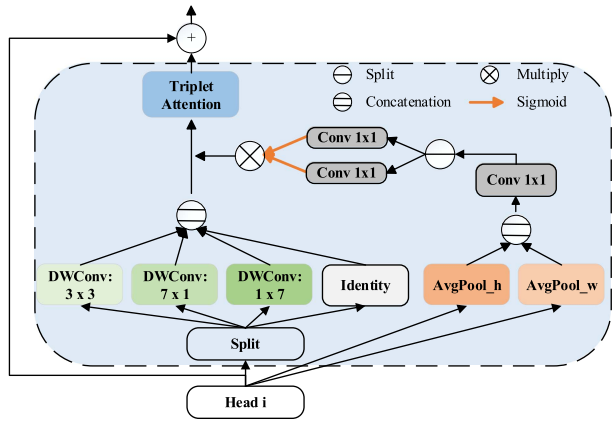

Fig. 4. Illustration of the proposed MATM.

gain better multiscale information and a faster learning speed, we introduce the attention mechanism [42], [43] to multiply the multiscale output features with the attention weight distribution. The feature map with localized information is finally obtained after two branches of feature extraction.

Specifically, for the spatial branch as shown in Fig. 2(a), we extract the multiscale features through convolutional layers with 3×3, 5×5, and 7×7. After the channel dimensions are merged, we apply a BN and a ReLU function to the network to provide stability and nonlinear features. Next, we use a 1×1 convolution layer followed by a BN and a ReLU function to reduce the number of channels and produce a multiscale feature map M_{spa}, which are then multiplied point-by-point with the resulting weights of the spatial attentional distributions. For the purpose of achieving faster convergence and more stable output, we also introduce a residual connection. The output of the spatial branch can be represented as follows:

$$F_{\text{spa}} = \sigma\left(f^{1\times1}\left(\text{Pool}\left(F\right)\right)\right) \odot M_{\text{spa}} + F \tag{2}$$

where σ stands for the sigmoid function, Pool contains the maximum pooling and average pooling of spatial dimensions, $\odot$ denotes the Hadamard product of matrices, and $f^{1\times1}$ is a 1x1 2-D convolutional layer.

For the spectral branch, as shown in Fig. 2(b), we execute feature extraction by the 1×1 group convolutions with three distinct scales of 2, 4, and 8. We follow the same dimensionality reduction procedure as for the spatial branch to obtain the M_{spe}, which is subsequently multiplied by the weight of the spectral branch's attention distribution. The output of the spectral branch can be summarized as follows:

$$F_{\text{spe}} = \sigma\left(f^{1\times1}\left(\delta\left(f^{1\times1}\left(\text{Pool}\left(F\right)\right)\right)\right)\right) \odot M_{\text{spe}} + F \tag{3}$$

δ is the ReLU activation function.

2) Spatial Adaptive Weights Module: The features extracted through the two branches of the PS2A need to be effectively fused. In the majority of previous methods, the spatial and spectral two-branch classification models frequently incorporate a linear layer or convolutional layer to simply perform the fusion and reduce the spectral dimensions [22], [59]. These methods render it easy to eliminate the spectral information while further extracting the spatial features. Consequently, we designed the SAWM as illustrated in Fig. 3 to ensure that no useful information is lost.

To get a more stable feature distribution, the PS2A feature maps are first normalized by BN following channel dimension merging. On the one hand, the local information in both branches is fused with deep convolution, and then the local signal is activated using a ReLU function. On the other hand, to efficiently extract the spectral information in specific dimensions, we use the spatial attention mechanism, but we construct the feature maps with half of the original channel dimensions following the linear layer. Finally, the outputs of the two branches are multiplied and fed into a linear layer for final information fusion. The SAWM can be expressed as follows:

$$T_1 = \delta\left(\text{DW}\left(\text{BN}\left(\left[F_{\text{spe}}; F_{\text{spa}}\right]\right)\right)\right) \tag{4}$$

$$T_2 = \sigma\left(f\left(G_{\text{avg}}\left(\text{BN}\left(\left[F_{\text{spe}}; F_{\text{spa}}\right]\right)\right)\right)\right) \tag{5}$$

$$F_{\text{SAWM}} = f(T_1 \odot T_2) \tag{6}$$

[;] denotes combining the outputs F_{spe} and F_{spa} of the two branches of the PS2A in the channel dimension, DW denotes deep convolution with the number of channels halved, G_{avg} denotes doing average pooling and maximum pooling in the spatial dimension, f is a linear layer.

B. GMCD Embedding

To capture the local properties between consecutive spectra at different locations and to preserve the invariance of the spatial structure of the original HSI blocks, we adopt a GMCD embedding strategy to group the feature sequentially along the spectral

dimensions as illustrated in Fig. 1(b). In detail, the F_{SAWM} is split into h groups along the channel dimension

$$F_{\mathrm{AFM}} = \{F_1, F_2, \ldots, F_k, \ldots, F_h\} \tag{7}$$

where h represents the number of attention heads, k is the sequence number for each group. F_k is the feature entered into each MATM, which reconstructs the barred or blocky morphological features through asymmetric convolution [60] and captures the spatial and channel cross-dimensional interaction through triplet attention [45]. Then, the output feature map of MATM goes through a global average pooling layer, after which it is reshaped to the shape of the tokens input by the transformer. After that, the tokens of different groups are merged in the channel dimension and then enter a linear layer to initially fuse the information of interacting different groups.

Specifically, as shown in Fig. 4, for each head F_k, MATM contains three different scales of asymmetric deep convolutional branches, a linear mapping layer, and attention branches in both horizontal and vertical directions. For the F_k of each head input, we initially split it into four parallel branches along the channel dimension, i.e., the small square kernel, two orthogonal strip kernels, and a constant mapping,

$$F_k^{hw}, F_k^{w}, F_k^{h}, F_k^{id} = \mathrm{Split}\left(F_k\right). \tag{8}$$

With reference to the default settings in [60], the number of each group in our $F_k^{hw}, F_k^{w}, F_k^{h}$ is 1/8 of the total number of F_k channels, i.e., $B \times 1/8$. F_k^{id} is a linear mapping layer containing $\frac{5}{8}B$. The split inputs are then embedded into the four parallel branches

$$\hat{F}_k^{hw} = \mathrm{DWConv}_{k_s \times k_s}\left(F_k^{hw}\right) \tag{9}$$

$$\hat{F}_k^{w} = \mathrm{DWConv}_{1 \times k_b}\left(F_k^{w}\right) \tag{10}$$

$$\hat{F}_k^{h} = \mathrm{DWConv}_{k_b \times 1}\left(F_k^{h}\right) \tag{11}$$

$$\hat{F}_k^{id} = F_k^{id} \tag{12}$$

where k_s denotes a square convolutional kernel size with smaller convolutional kernels, defaulting to 3, and k_b denotes an asymmetric banded convolutional kernel size defaulting to 7. Finally, the features of each branch are merged

$$\hat{F}_k = \mathrm{Concat}\left(\hat{F}_k^{hw}, \hat{F}_k^{w}, \hat{F}_k^{h}, \hat{F}_k^{id}\right). \tag{13}$$

It is common knowledge that location information plays a significant role in producing spatially selective attention maps [37], [44]. Therefore, we conduct a global average pooling of the feature maps in both vertical, horizontal, and orientations throughout the spatial dimension. Each element of the cth channel in F_k in the horizontal direction is given by

$$z_c^h\ (h) = \frac{1}{W}\sum_{i=0}^{W-1} f_c\left(h, i\right). \tag{14}$$

Similarly, each element of the cth channel in F_k in the vertical direction is given by

$$z_c^w\ (w) = \frac{1}{H}\sum_{j=0}^{H-1} f_c\left(j, w\right). \tag{15}$$

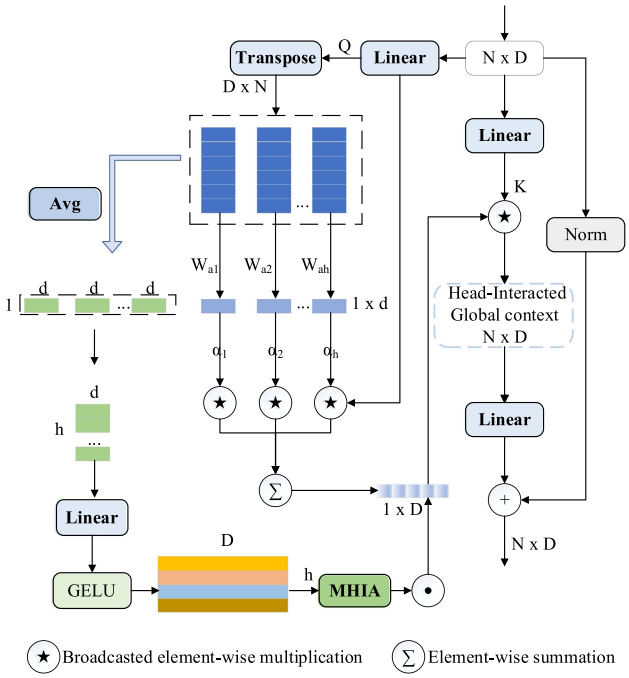

Fig. 5. Illustration of the MHIASA.

The resulting z^h and z^w are merged and then fed into a shared 1x1 2-D convolutional layer, this procedure can be represented as

$$M = \delta\left(f^{1\times1}\left[z^h; z^w\right]\right). \tag{16}$$

The resulting feature map M is an intermediate feature map containing spatial information in both directions. In the spatial dimension, M is then divided into two independent tensors M_h and M_w, which proceed into the convolutional layer and sigmoid in two separate directions. Then two attention distributions W_h and W_w are generated. The following is a description of this process:

$$W_h = \sigma\left(f^{1\times1}\left(M_h\right)\right) \tag{17}$$

$$W_w = \sigma\left(f^{1\times1}\left(M_w\right)\right). \tag{18}$$

The obtained attention weights are multiplied element-by-element with the aggregated features $\hat{F}_k$ and then enter the triplet attention module after the norm operation.

$$\hat{F}_k^{wh} = \mathrm{Norm}\left(\hat{F}_k \odot W_w \odot W_h\right). \tag{19}$$

Afterward $\widehat{F}_k^{wh}$ is fed into the triplet attention module. The output feature map $\hat{F}_k^t$ of the triplet attention module and the feature map F_k of the input MATM are linked by residuals to reduce overfitting.

C. Multihead Interacted Additive Self-Attention

The learning ability of the channel representations can be increased by including the interaction between multiple attentional heads, which can fuse the various representations into the overall representation of the object. Thus, an MHIASA module in Fig. 5 is presented. Specifically, the embedding matrix X obtained by GMCD is linearly projected into a query vector

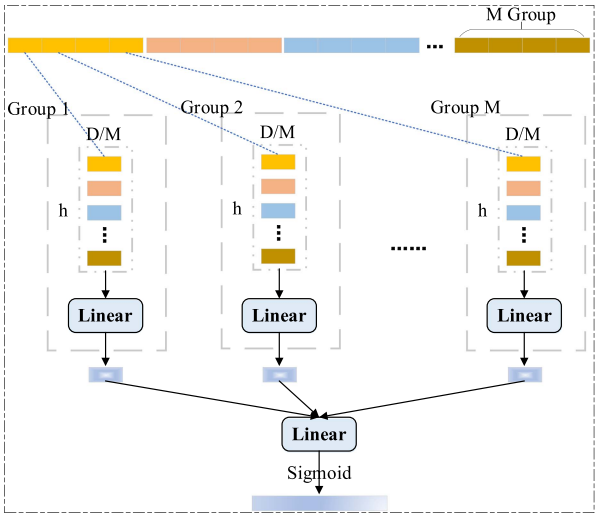

Fig. 6. Structure of the MHIA module.

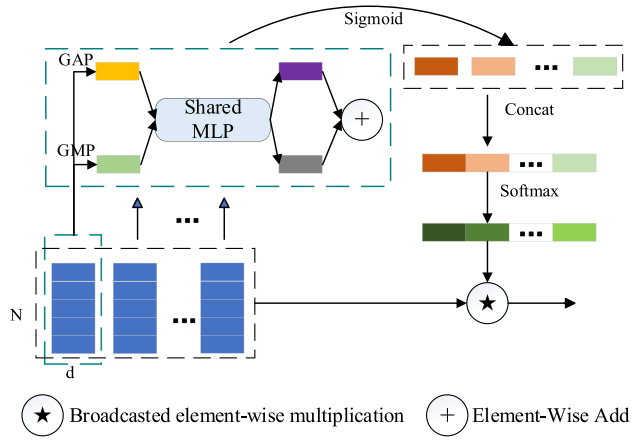

Fig. 7. Structure of AMHCat.

Q and a key vector K with two weight matrices W_q, W_k, where $Q, K \in R^{N\times D}, W_q$, $W_k \in R^{D\times D}, N$ is the length of the token sequence, and D is the dimensionality of the embedding vector. To extract the global information relationship of the spectral dimension effectively, we transpose Q. Next, the Q^T is multiplied with the learnable parameter vector $w_a \in R^N$ to learn the attention weight distribution. This is followed by a softmax operation to generate the global attention query vector $\alpha \in R^D$

$$\alpha = \frac{\exp\left(Q^T \cdot \frac{w_a}{\sqrt{N}}\right)}{\sum_{j=1}^{D}\exp\left(Q^T \cdot \frac{w_a}{\sqrt{N}}\right)}. \tag{20}$$

The query matrix then generates a single global query vector $q \in R^D$ based on the learned attention weight distribution

$$q = \sum_{i=1}^{N} \alpha_i * Q_i \,. \tag{21}$$

The multihead interacted aggregation (MHIA) module in Fig. 6 is the central component of MHIASA, and it processes $Q^T \in R^{D\times N}$ in the following way. Let the number of patch tokens be N and the number of attentional heads be h. Then each D dimensional token will be reshaped into h parts, and by averaging each part, we obtain h vectors of size $1 \times d$. These h vectors are projected to the D-dimension space after passing through linear and GELU to get $H \in R^{D\times h}$ before entering the MHIA. Specifically, MHIA reorganizes the arrangement of the h head vectors, and we select D/M channels from each head to construct a group and then fuse the h head vectors into one at each group through a linear mapping, ensuring feature aggregation between each attention head. Subsequently, we use a linear layer to perform cross-group information fusion of intragroup–intergroup patterns to achieve efficient aggregation results and obtain the final result $M_{\text{MHIA}} \in R^D$, where $G_i \in R^{\frac{D}{M}\times h}$

$$M_{\text{MHIA}} = \sigma\left(W_{\text{inter}}\left([G_1, G_2, \ldots, G_M]\right)\right) \tag{22}$$

$$G_i = W_{\text{intra}}\left(\left[H_1^i, H_2^i, \ldots H_h^i\right]\right)\ i \in [1, M] \tag{23}$$

where W_{inter} and W_{intra} are the weight matrices of the linear layer. Then, the output M_{MHIA} of the MHIA is multiplied with the q to obtain a global query vector $\hat{q}$ with attentional head interaction information. Next, $\hat{q}$ and K are encoded with broadcast multiplication of the information to produce a head-interacted global context vector. This vector extracts information from each token and has the flexibility to obtain the relationship between each input sequence and also contains the interaction information between each attention head. We have expanded these capabilities simply by increasing the linear complexity. The entire process is shown in Fig. 5. The final output of the MHIASA $X_{\text{out}} \in R^{N\times D}$ can be represented as

$$X_{\text{out}} = \hat{X} + f\left(K * \hat{q}\right). \tag{24}$$

$\hat{X}$ represents the original vector after normalization. The output X_{out} goes to the MHSA module for further feature extraction.

D. Adaptive Multihead Catenation

The self-attention mechanism in a traditional transformer [48] utilizes each attention head to extract feature information in various directions, but it ignores the attention weights from each head. The output for each head $Z_i \in R^{d\times N}$ in AMHCat is

$$\hat{Z}_i = \sigma\left(\text{MLP}\left(\text{Pool}\left(Z_i\right)\right)\right)\ i \in [1, h]. \tag{25}$$

Consequently, we augment MHSA with the CBAM [43], which obtains the attention weights of features in various orientations as shown in Fig. 7. Finally, the acquired attentional weights are concatenated and normalized by using the softmax function. Accordingly, the final output $Z \in R^{D\times N}$ of AMHCat is

$$S = \text{softmax}\left(\text{Concat}\left(\hat{Z}_1, \ldots, \hat{Z}_h\right)\right) \tag{26}$$

$$Z = S * \text{Concat}\left(Z_1, \ldots, Z_h\right). \tag{27}$$

IV. Experiments and Results

This section provides a detailed description of the pertinent aspects of experiments, including data description, experimental

TABLE I
LAND-COVER CATEGORY AND NUMBER OF TRAINING, VALIDATION, AND TESTING SAMPLES ON THE IP DATASET

No.	Name	Train.	Val.	Test.
1	Alfalfa	2	3	41
2	Corn-notill	71	72	1285
3	Corn-mintill	42	41	747
4	Corn	12	12	213
5	Grass-pasture	24	24	435
6	Grass-trees	37	36	657
7	Grass-pasture-mowed	2	1	25
8	Hay-windrowed	24	24	430
9	Oats	1	1	18
10	Soybean-notill	49	48	875
11	Soybean-mintill	122	123	2210
12	Soybean-clean	30	29	534
13	Wheat	10	10	185
14	Woods	63	63	1139
15	Buildings-Grass-Trees-Drives	19	20	347
16	Stone-Steel-Towers	4	5	84
	Total	512	512	9225

TABLE II
LAND-COVER CATEGORY AND NUMBERS OF TRAINING, VALIDATION, AND TESTING SAMPLES ON THE PU DATASET

No.	Name	Train.	Val.	Test.
1	Asphalt	66	66	6499
2	Meadows	186	187	18 276
3	Gravel	21	21	2057
4	Trees	30	31	3003
5	Painted metal sheets	14	13	1318
6	Bare Soil	50	50	4929
7	Bitumen	13	14	1303
8	Self-Blocking Bricks	37	37	3608
9	Shadows	10	9	928
	Total	427	428	41 921

TABLE III
LAND-COVER CATEGORY AND NUMBERS OF TRAINING, VALIDATION, AND TESTING SAMPLES ON THE WHU-HC DATASET

No.	Name	Train.	Val.	Test.
1	Strawberry	223	224	44 288
2	Cowpea	113	114	22 526
3	Soybean	51	52	10 184
4	Sorghum	27	27	5299
5	Water spinach	6	6	1188
6	Watermelon	23	22	4488
7	Greens	30	29	5844
8	Trees	90	90	17 798
9	Grass	47	48	9374
10	Red roof	52	53	10 411
11	Gray roof	84	85	16 742
12	Plastic	19	18	3642
13	Bare soil	46	45	9025
14	Road	93	93	18 374
15	Bright object	6	5	1125
16	Water	377	377	74 647
	Total	1287	1288	254 955

TABLE IV
LAND-COVER CATEGORY AND NUMBERS OF TRAINING, VALIDATION, AND TESTING SAMPLES ON THE WHU-LK DATASET

No.	Name	Train.	Val.	Test.
1	Corn	172	173	34 166
2	Cotton	42	42	8290
3	Sesame	15	15	3001
4	Broad-leaf soybean	316	316	62 580
5	Narrow-leaf soybean	21	21	4109
6	Rice	59	60	11 735
7	Water	335	335	66 386
8	Roads and houses	36	35	7053
9	Mixed weed	26	26	5177
	Total	1022	1023	202 497

setup, parameter analysis, classification results and analysis, and ablation studies.

A. Data Description

In this work, four publicly available HSI datasets are employed to evaluate the effectiveness of our proposed method, including Indian Pine dataset (IP), Pavia University dataset (PU), WHU-Hi-HanChuan dataset (WHU-HC), WHU-Hi-LongKou dataset (WHU- LK). We only used 5%, 1%, 0.5%, and 0.5% data on the four datasets as training sets, respectively. Tables I–IV show the land cover categories of the used datasets and the number of samples in the training, validation, and test sets.

1) *IP Dataset:* This dataset is the earliest dataset used for HSI classification and was created by the airborne visual infrared imaging spectrometer (AVIRIS) in 1992 by imaging a patch of Indian Pine trees in Indiana, USA. The AVIRIS imaging spectrometer has a wavelength range of 0.4–2.5 μm and was designed to image features continuously in 220 consecutive bands. However, since bands 104–108, 150–163, and 220 are not reflective of water, we generally use the 200 bands that remain after these 20 bands have been eliminated from the study. The IP consists of 16 land-cover types and 145 $\times$ 145 pixels.
2) *PU Dataset:* The PU dataset was acquired by the Reflectance Optical System Imaging Spectrometer in 2001. The spectral wavelength range is 380–860 nm with 115 spectral bands. After removing the noise bands, 103 usable bands were used for the study. The dataset contains 610 $\times$ 340 pixels and 9 land cover types.
3) *WHU-HC Dataset:* The WHU-HC dataset was acquired on 17 June 2016, in Hanchuan City, Hubei Province, China, using a Headwall Nano-Hyperspec imaging sensor with a focal length of 17 mm powered by a Leica Aibot X6 UAV V1 platform. The study area was an agricultural scene in the urban–rural area. The spatial resolution is about 0.109 m, with 274 bands and a spectral wavelength range of 400–1000 nm. The image contains 1217 $\times$ 303 pixels and 16 land cover types.
4) *WHU-LK Dataset:* The WHU- LK dataset was acquired on 17 July 2018, in Longkou Town, Hubei Province, and the study area is a simple agricultural scene containing six crop species. The spatial resolution is about 0.463 m, and the spectral wavelength range is 400–1000 nm. The dataset contains 550 $\times$ 400 pixels in 270 bands and 9 land cover types.

TABLE V
CLASSIFICATION RESULTS OBTAINED BY DIFFERENT METHODS FOR INDIAN PINES DATASET

Class	DFFN	HybridSN	SF	GAHT	SSFTT	BS2T	morphFormer	MHIAIFormer
1	54.63±16.37	57.56±24.66	23.41±13.85	18.54±16.73	**89.27±6.28**	38.54±33.67	73.66±13.31	79.03±18.80
2	85.87±2.56	77.60±7.30	76.70±3.59	91.63±4.57	**95.16±2.94**	89.91±3.15	92.45±3.38	94.29±3.17
3	91.51±3.94	65.62±9.12	76.52±5.80	91.14±1.91	90.98±6.38	86.24±13.00	94.32±2.62	**94.54±3.35**
4	78.69±17.15	75.87±9.79	82.54±4.86	88.54±2.65	**97.75±1.86**	83.19±19.10	91.74±3.55	95.59±3.58
5	91.26±2.67	87.68±5.10	84.83±5.40	81.01±18.94	90.90±3.91	90.21±6.94	90.48±5.35	**91.03±3.27**
6	92.33±10.93	96.59±1.71	95.95±1.45	98.96±0.50	**99.54±0.54**	98.08±1.26	98.51±1.32	99.33±0.77
7	54.40±38.08	43.20±18.66	60.00±25.17	24.80±32.44	**93.60±10.91**	56.80±46.54	73.60±18.52	88.00±11.87
8	96.23±4.06	95.44±4.14	97.95±1.76	**99.95±0.09**	99.44±1.12	99.30±1.40	99.53±0.64	98.79±0.58
9	30.00±26.20	38.89±22.50	38.89±8.61	0.00±0.00	58.89±25.48	3.33±6.67	54.44±17.00	**84.44±13.79**
10	80.64±10.66	82.74±6.13	76.69±0.48	90.72±3.06	**95.11±2.08**	80.94±14.29	94.47±1.66	94.06±2.35
11	89.35±6.42	90.72±3.16	86.82±2.83	95.16±1.04	95.11±1.83	**97.52±1.77**	96.05±0.83	96.74±0.57
12	68.69±9.96	63.11±10.14	62.36±5.97	90.64±4.70	84.64±5.72	85.47±12.80	88.13±6.73	**96.18±1.14**
13	95.03±4.18	99.14±0.88	95.46±2.62	95.57±3.37	**99.89±0.22**	95.14±3.87	98.81±0.86	98.60±0.81
14	96.93±1.57	98.14±1.06	95.24±1.36	96.91±2.84	98.65±1.97	98.60±1.13	**99.26±0.53**	99.07±0.59
15	79.94±8.55	85.59±3.77	80.58±5.55	87.09±5.68	93.26±5.58	82.31±17.74	89.97±4.46	**94.41±3.06**
16	93.10±6.09	**99.52±0.58**	91.19±5.86	69.05±13.04	92.62±4.67	82.38±18.48	91.43±5.90	90.71±5.19
OA (%)	87.76±2.10	85.25±2.40	83.75±0.35	92.30±1.60	94.90±0.56	91.54±2.17	94.72±0.57	**95.97±0.76**
AA (%)	79.91±2.05	78.59±3.00	76.57±0.99	76.23±3.96	92.18±1.67	79.25±6.60	89.18±1.38	**93.43±1.88**
$k \times 100$	86.03±2.45	83.08±2.79	81.45±0.40	91.20±1.83	94.19±0.63	90.30±2.55	93.98±0.65	**95.41±0.87**

TABLE VI
CLASSIFICATION RESULTS OBTAINED BY DIFFERENT METHODS FOR PU DATASET

Class	DFFN	HybridSN	SF	GAHT	SSFTT	BS2T	morphFormer	MHIAIFormer
1	95.64±1.50	95.39±1.72	89.37±3.38	95.86±3.57	**98.31±0.71**	98.99±1.13	97.56±0.56	98.14±1.07
2	98.44±0.66	96.68±1.22	97.13±1.12	98.69±1.42	99.78±0.13	99.75±0.23	99.49±0.27	**99.85±0.06**
3	79.79±15.91	74.27±10.69	64.61±5.26	84.70±15.97	90.84±3.89	85.08±8.97	90.50±5.29	**96.16±2.17**
4	93.90±2.06	97.00±0.71	90.73±2.61	95.65±0.58	**98.33±0.95**	92.02±5.64	93.61±2.26	96.48±2.10
5	99.44±0.60	**100.00±0.00**	99.64±0.27	99.98±0.03	99.54±0.32	99.89±0.10	99.79±0.19	**100.00±0.00**
6	95.83±2.76	82.26±6.37	89.08±5.38	95.73±5.97	98.38±0.81	99.26±0.52	**99.63±0.18**	98.88±0.65
7	84.10±8.12	80.92±11.61	40.21±13.03	89.98±5.26	90.39±6.40	88.24±12.86	90.79±3.63	**97.18±2.15**
8	87.82±6.96	90.96±2.72	78.24±2.89	93.97±2.6	95.55±2.58	96.33±3.39	95.93±1.77	**97.06±1.30**
9	96.90±1.07	99.68±0.33	91.62±2.51	95.11±2.43	**98.06±1.29**	94.81±2.34	95.50±2.31	97.85±1.40
OA (%)	95.10±1.09	92.89±0.94	89.49±0.76	96.29±2.90	98.14±0.20	97.54±0.48	97.69±0.50	**98.68±0.09**
AA (%)	92.43±1.65	90.80±2.28	82.29±1.69	94.41±3.80	96.58±0.79	94.93±1.01	95.87±1.13	**97.95±0.29**
$k \times 100$	93.50±1.44	90.55±1.28	86.01±1.04	95.08±3.84	97.54±0.27	96.73±0.65	96.93±0.66	**98.26±0.13**

TABLE VII
CLASSIFICATION RESULTS OBTAINED BY DIFFERENT METHODS FOR WHU-HC DATASET

Class	DFFN	HybridSN	SF	GAHT	SSFTT	BS2T	morphFormer	MHIAIFormer
1	97.25±1.03	94.29±1.90	94.97±1.50	98.36±0.70	96.32±0.93	96.64±3.47	96.60±1.71	**98.54±0.42**
2	83.94±7.11	82.49±3.49	78.44±1.65	87.05±1.59	90.24±2.37	89.50±3.44	93.82±1.35	**93.57±2.72**
3	76.89±7.14	76.08±21.13	85.42±7.21	58.44±9.52	86.00±7.19	**92.87±3.58**	91.78±4.23	91.86±4.29
4	76.30±13.28	92.29±1.42	89.25±5.30	51.09±35.03	89.86±12.17	95.52±2.89	**95.60±1.26**	91.58±5.10
5	42.02±20.96	50.61±9.89	35.64±16.13	2.53±3.91	66.99±21.21	72.46±26.79	63.10±17.30	**85.15±10.23**
6	11.57±9.38	16.41±13.48	18.88±6.69	2.83±3.97	55.35±7.33	**64.12±15.43**	49.06±7.00	59.09±12.87
7	69.10±17.63	77.53±10.35	68.04±8.74	70.79±7.22	83.44±7.47	**86.67±6.99**	81.91±5.71	83.79±11.20
8	59.00±16.56	66.55±8.29	71.47±3.47	64.84±3.61	81.32±7.65	**86.94±4.15**	80.33±2.87	83.22±1.16
9	59.54±7.70	48.81±10.05	55.08±7.80	37.32±20.66	**85.34±3.95**	82.75±17.15	82.10±4.77	81.26±2.93
10	83.92±11.66	91.02±3.10	85.84±5.09	69.88±33.78	94.64±1.79	94.26±2.95	93.48±5.66	**95.37±2.00**
11	83.60±5.33	93.39±2.52	82.35±6.32	74.88±12.85	93.35±5.54	**96.89±2.04**	91.92±7.71	90.41±9.37
12	10.64±8.86	27.62±10.61	28.40±6.70	3.13±5.41	71.47±8.53	53.79±26.55	52.78±18.31	**77.97±7.81**
13	34.80±8.40	48.69±4.14	52.79±2.93	33.27±15.67	66.85±3.05	**75.18±8.66**	72.65±6.17	69.42±7.71
14	82.67±3.71	82.01±3.72	88.97±2.42	83.61±5.21	86.39±2.76	89.85±3.92	89.92±1.88	**92.04±3.73**
15	51.43±10.86	**76.59±9.03**	36.21±15.01	0.52±1.03	74.26±7.89	63.68±15.60	70.67±13.25	75.41±8.94
16	99.01±0.82	98.85±0.62	97.74±1.56	98.88±0.59	99.04±0.60	99.25±0.38	99.68±0.16	**99.74±0.16**
OA (%)	82.73±1.84	84.52±1.31	84.14±0.96	79.73±3.79	90.95±0.97	92.41±1.47	91.54±0.85	**92.68±1.01**
AA (%)	63.86±3.90	70.20±3.32	66.84±1.64	52.34±6.98	82.55±2.52	83.77±3.81	81.59±2.27	**85.53±1.92**
$k \times 100$	79.69±2.15	81.86±1.56	81.42±1.10	76.03±4.60	89.40±1.13	91.10±1.73	90.09±1.00	**91.42±1.20**

B. Experimental Setup

1) Evaluation Metrics: Three frequently used evaluation metrics, overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa), were selected to quantitatively assess the classification performance of all methods. To reduce the effect of experimental randomness, all models were run five times, and the average results and standard deviations were recorded. In addition, we visualize the classification color plots to qualitatively compare the results obtained by different methods.

TABLE VIII
CLASSIFICATION RESULTS OBTAINED BY DIFFERENT METHODS FOR WHU-LK DATASET

Class	DFFN	HybridSN	SF	GAHT	SSFTT	BS2T	morphFormer	MHIAIFormer
1	99.73±0.09	99.78±0.10	99.69±0.13	99.91±0.07	99.83±0.09	**99.93±0.06**	99.74±0.10	99.78±0.09
2	92.83±4.48	89.18±11.29	91.32±2.67	96.71±1.30	98.95±0.55	98.64±1.40	97.99±1.39	**99.10±0.61**
3	91.54±4.18	90.76±3.83	94.38±4.69	93.84±4.61	**97.45±1.83**	75.77±24.01	95.64±3.17	96.91±2.07
4	98.13±1.20	98.37±0.93	98.56±0.34	99.27±0.46	99.55±0.19	99.72±0.15	99.65±0.23	**99.76±0.09**
5	64.64±23.28	74.10±20.69	80.31±5.95	87.91±4.15	95.14±1.23	92.13±4.21	89.07±4.66	**95.16±2.81**
6	99.16±0.74	98.46±1.85	97.27±1.20	99.52±0.34	99.56±0.61	99.20±0.59	98.80±0.56	**99.77±0.21**
7	99.94±0.03	**99.99±0.01**	99.98±0.02	99.91±0.08	99.93±0.04	99.86±0.15	99.96±0.04	99.96±0.04
8	91.47±5.93	94.20±2.23	89.88±4.02	94.04±4.21	95.05±1.83	93.86±4.10	93.25±4.11	**96.95±0.71**
9	84.94±7.22	87.32±5.20	88.28±2.80	91.08±3.08	95.00±3.93	88.00±5.99	90.64±3.88	**96.92±0.87**
OA (%)	97.49±0.64	97.74±0.60	97.85±0.19	98.79±0.31	99.31±0.08	98.72±0.38	98.92±0.16	**99.49±0.07**
AA (%)	91.38±1.88	92.46±2.16	93.30±1.06	95.80±1.13	97.83±0.38	94.13±2.68	96.08±0.85	**98.26±0.30**
$k \times 100$	96.70±0.83	97.02±0.79	97.17±0.25	98.41±0.41	99.09±0.11	98.31±0.50	98.58±0.21	**99.34±0.10**

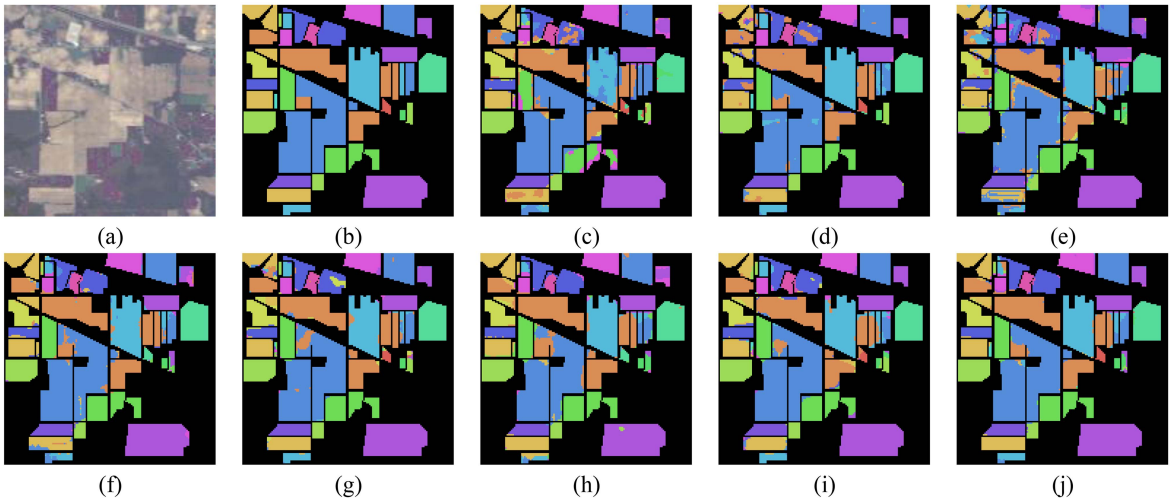

Fig. 8. Classification maps using different classification methods on the IP dataset. The color labels correspond to Table I. (a) False-color image. (b) Ground-truth map. (c) DFFN. (d) HybridSN. (e) SF. (f) GAHT. (g) SSFTT. (h) BS2T. (i) morphFormer. (j) Proposed MHIAIFormer.

2) Comparison Methods: In order to validate the effectiveness of the proposed algorithms, we selected several of the most representative HSI classification methods for comparison, including pure CNN models: DFFN [19], HybridSN [20]; pure transformer models: SF [31], GAHT [32]; hybrid models of CNN and transformer: SSFTT [38], BS2T [39], and morphFormer [40].

1) DFFN is a CNN-based network that deepens the network depth by superimposing 2-D convolutional blocks and obtains different hierarchical information by introducing residual structures to construct shallow, middle, and deep layers. Then, a feature fusion mechanism is introduced to utilize complementarity and relevance between different levels of information for HSI classification.
2) The HybridSN extracts features using a combination of 2-D and 3-D convolutions; it has two dropout layers, three full connection layers, three 3-D convolution layers, and one 2-D convolution layer.
3) SF proposes a new backbone network based on the transformer that is capable of learning spectral localization information in the form of embeddings grouped in adjacent bands. Then the application of skip connection in the transformer is innovated, and a cross-layer adaptive fusion module is developed, which effectively combines the shallow and deep features.
4) SSFTT is a transformer-based method. SSFTT extracted shallow spatial-spectral features using 2-D and 3-D convolutional layers. Subsequently, the features are converted into tokens with high-level semantic information using a Gaussian-weighted feature tokenizer. These tokens are then fed into the transformer encoder to facilitate additional feature learning.
5) GAHT is a transformer-based method that proposes an embedding method for grouping along the channel dimensions. Group convolution is applied to segment patches by pixels. GAHT adopts a hierarchical structure overall and minimizes resource usage by gradually reducing the number of channels and mining the useful features.
6) BS2T is a transformer-based method that contains three stages. The first stage extracts local features by 3-D convolution. The second stage introduces spectral information and contextual spatial location information into MHSA. The last stage fuses the feature information of the spectral and spatial branches for classification.

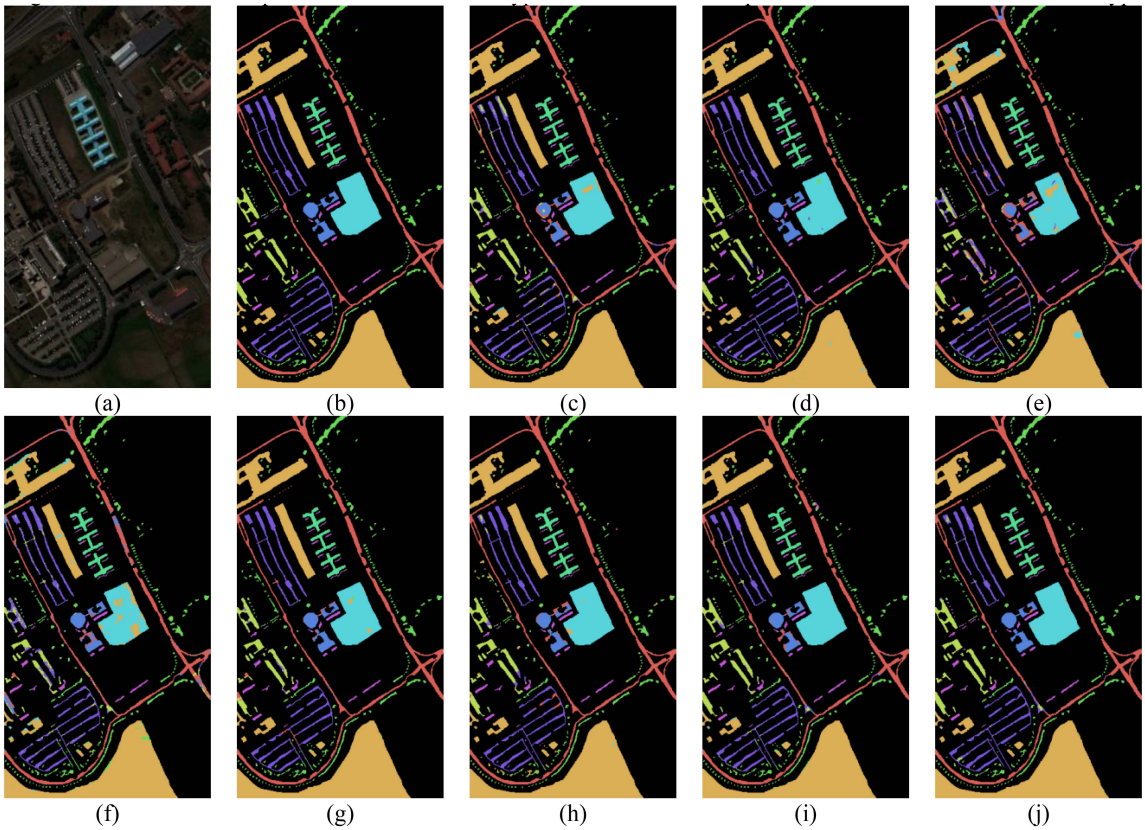

Fig. 9. Classification maps using different classification methods on the PU dataset. The color labels correspond to Table II. (a) False-color image. (b) Ground-truth map. (c) DFFN. (d) HybridSN. (e) SF. (f) GAHT. (g) SSFTT. (h) BS2T. (i) morphFormer. (j) Proposed MHIAIFormer.

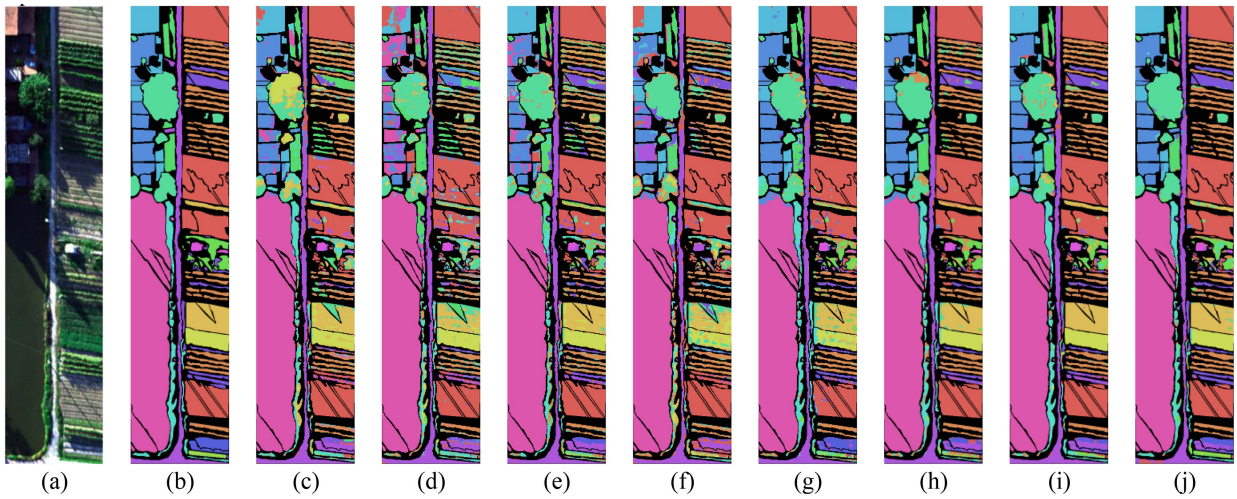

Fig. 10. Classification maps using different classification methods on the WHU-HC dataset. The color labels correspond to Table III. (a) False-color image. (b) Ground-truth map. (c) DFFN. (d) HybridSN. (e) SF. (f) GAHT. (g) SSFTT. (h) BS2T. (i) morphFormer. (j) Proposed MHIAIFormer.

7) morphFormer is a new transformer that introduces morphology and employs spectral and spatial morphological convolution operations with dilation and erosion operators in conjunction with an attention mechanism to improve the interaction between the structure and shape information of HSI and CLS tokens.

3) Implementation Details: To ensure that the comparison experiments are notarized and comparable, all models are run on the DL framework PyTorch 2.2.0. Batch size and epoch are set to 64 and 100, respectively, to update the training parameters of the models. The optimizer and learning rates are the same as the most suitable parameters in the articles to guarantee the models'

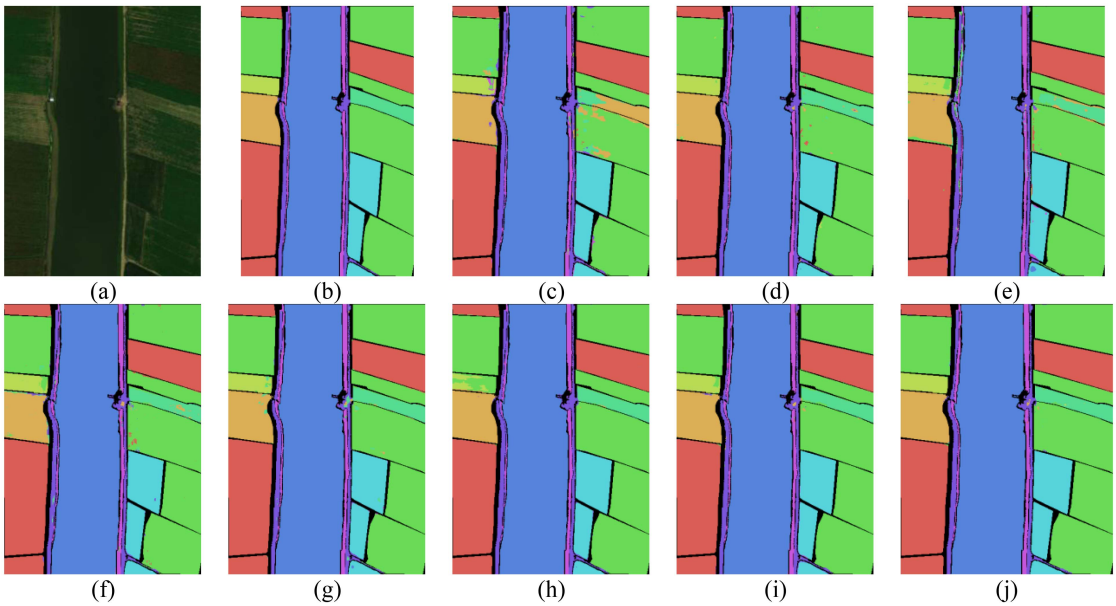

Fig. 11. Classification maps using different classification methods on the WHU-LK dataset. The color labels correspond to Table IV. (a) False-color image. (b) Ground-truth map. (c) DFFN. (d) HybridSN. (e) SF. (f) GAHT. (g) SSFTT. (h) BS2T. (i) morphFormer. (j) Proposed MHIAIFormer.

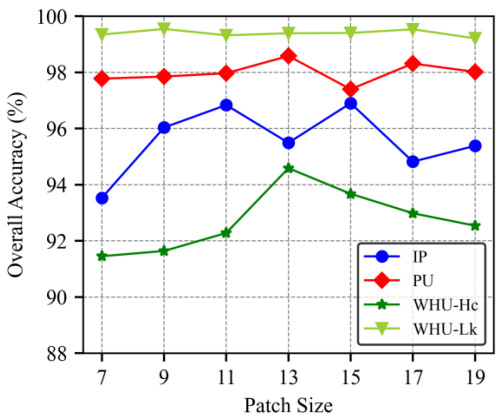

Fig. 12. OA (%) with different spatial sizes on the four datasets.

best performance, which is suggested in other publications. For our proposed method, we use the Adam optimizer, where β_1 is set to 0.9 and β_2 is set to 0.999. For simplicity, the learning rate is a constant of 0.001. We used an NVIDIA GeForce RTX 2080 SUPER graphics processing unit with 8 GB of memory to accelerate the experiments.

C. Classification Results and Analysis

We conducted extensive experiments on IP, PU, WHU-HC, and WHU-LK datasets and compared them on OA, AA, and Kappa metrics. As shown in Tables V–VIII, bold values represent the best classification results, the results of our proposed model on OA, AA, and Kappa are superior to other methods. We can draw the following conclusions from the results. First, DFFN and HybridSN, which are CNN-based models, usually produce worse results than transformer-based methods. In addition, transformer-based methods still lag behind our proposed method, although they show excellent potential for capturing long-distance spectral-spatial features. The main reason is that the 3D attributes of HSI cannot be effectively modeled with the same amount of data.

Each dataset is characterized by its region's shape, so the classification of different datasets has different challenges. We scrutinized the challenging classes that are prone to misclassification in each dataset. On the IP dataset, we use only 5% of the samples to train the model, and the biggest gap with other methods is found in classes with relatively small sample sizes, such as "Alfalfa" and "Oats." The distributions of these two categories are banded and irregular. The OA of DFFN, SF, GAHT, and BS2T in these two categories is low because of their lack of feature extraction capability. Our proposed model has the ability to obtain relatively homogeneous and higher results from limited training samples, with a 1.25% improvement over the 94.72% morphFormer of the best results in other methods. Similarly, on the WHU-HC dataset, this disparity is reflected in the results of the "Water spinach ," "Plastic," and "Bright object" categories with small sample sizes, demonstrating that our method has equally effective results on both large and small datasets. On the PU and WHU-LK datasets, our method also demonstrates extremely favorable results in most categories.

We chose the one with the highest OA effect for five experiments among the different methods to visualize each of the four datasets to qualitatively compare these models. The classification plots obtained for all methods are shown in Figs. 8, 9, 10, and 11. It can be seen that our proposed method possesses fewer noise points on the four datasets compared to the other methods. It can be realized that the classification results are closer to the truth map, further proving its superiority.

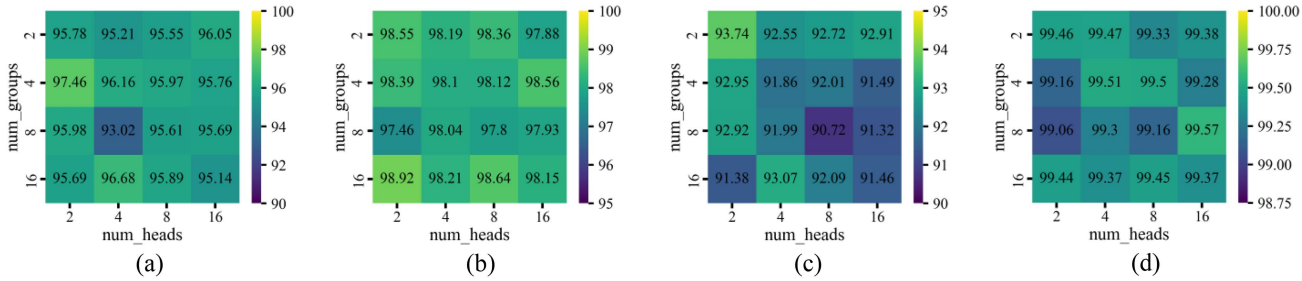

Fig. 13. Impact of the number of head and the number of groups on OA. (a) IP. (b) PU. (c) WHU-HC. (d) WHU-LK.

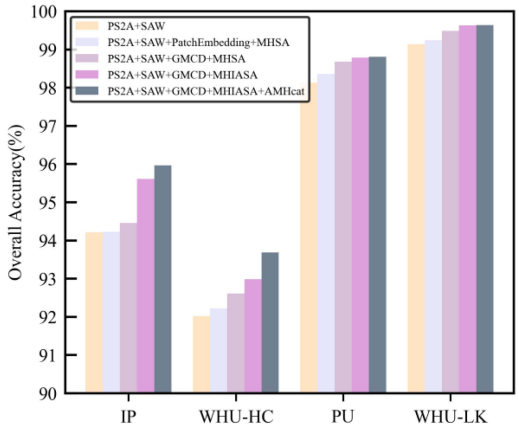

Fig. 14. Ablation study of the proposed modules.

V. Discussions

To thoroughly validate the effectiveness of our proposed method, we performed a series of experiments, which included the analysis of some important parameters as well as the effect of each part and robustness. In addition, we provide a detailed discussion of how other spatial-spectral pyramid-based models differ from the one proposed in the study.

A. Parameter Analysis

1) Patch Size: The size of the patch determines how much information is input to the model and has a very important role in the classification results. In this section, to evaluate the impact of patch size, we conducted experiments on four datasets with $\{7, 9, 11, 13, 15, 17, 19\}$. As shown in Fig. 12, the classification results are increasing with the increase in patch size. However, the classification accuracy does not always increase with the increase in patch size either. When the patch size is 13, the OA on the PU and WHU-HC datasets starts to decrease. When the patch size is 9 and 11, the OA on the WHU-LK and IP datasets starts to decrease, respectively. Taking into account the effects of accuracy and model computational loss, the patch size is set to 11×11 in all datasets and comparison models.

2) Numbers of Heads and Groups: The number of groups in GMCD, the number of attentional heads in MHSA, and the number of groups in MHIA have a crucial role in the classification results. We set the number of groups in GMCD equal to the number of attention heads in MHSA, and the range of both parameters is taken as $\{2, 4, 8, 16\}$. Fig. 13 illustrates the effect of these two parameters on the classification results of the four datasets. The optimal parameter combination is used to train the model on each of the four data points. The number of attentional heads is taken as 2 on the IP, PU, and WHU-HC datasets, and 16 on the WHU-LK dataset. The number of groups in the MHIA is taken as 4, 16, 2, and 8 on the IP, PU, WHU-HC, and WHU-LK datasets in that order.

B. Ablation Studies

1) Ablation Study of the Proposed Modules: This ablation analysis checks the validity of all the modules in the proposed MHIAIFormer, including PS2A, SAWM, GMCD, MHIASA, and AMHcat. We performed ablation experiments by adding different modules to the four datasets. The results are shown in Fig. 14. It can be seen that the worst results are produced when only PS2A and SAWM are used. After PatchEmbedding and MHSA are introduced, OA is initially improved. The OA is further improved after replacing PatchEmbdding with GMCD, which indicates that GMCD is effective in extracting features of different dimensions. After replacing MHSA with MHIASA, OA is significantly improved, proving that MHIASA can effectively improve the classification results while capturing different attention-head interaction features. The classification accuracy reaches its highest when AMHcat continues to be added.

2) Ablation Study of the Percentage of Training Samples: To demonstrate the robustness of the proposed model, we investigated the OA of MHIAIFormer and other comparative methods under different numbers of training samples. Specifically, for the IP dataset, 5%, 10%, 15%, 20%, and 25% of the training data were randomly selected in turn, and for the PU dataset, it was 0.25%, 5%, 1%, 2%, and 4%. For the WHU-HC and WHU-LK datasets, both are set to 0.25%, 0.5%, 0.75%, 1%, and 3%. As shown in Fig. 15, when the percentage is small, the proposed method shows greater enhancement results compared to other comparison methods. For example, in the PU dataset, when the sampling rate is set as 0.25%, the proposed method can achieve an OA of 91%, while for other DL-based method, such as SF, the OA is just nearly 76%. When the sampling rate is large, our method is still competitive among all the methods; however, the improvement is not

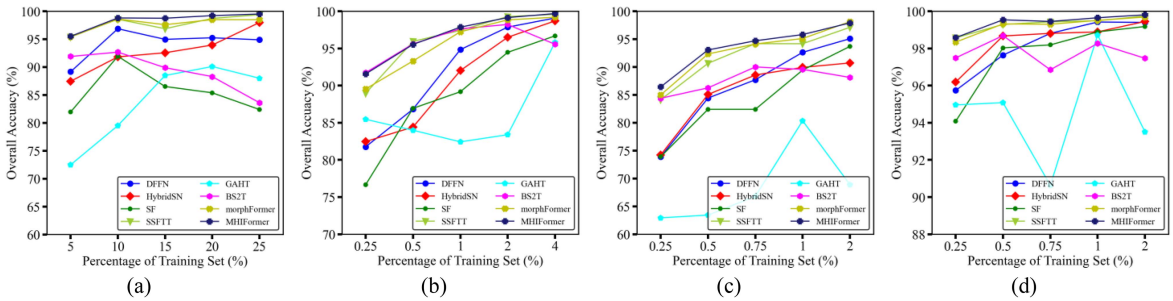

Fig. 15. Classification performance under different training samples. (a) IP. (b) PU. (c) WHU-HC. (d) WHU-LK.

such significant. In addition, on the performance of the four datasets, the proposed MHIAIFormer demonstrates comparable performance to SSFTT and morphFormer. It can be noticed that for the IP dataset SF and BS2T show overfitting with the increase of training data, and GAHT lacks robustness to the number of samples for PU, WHU-HC, and WHU-LK datasets. In comparison to other approaches, MHIAIFormer performs better, has superior OA, and demonstrates high sample sensitivity by changing gradually as the number of training samples increases.

C. Spatial-Spectral Pyramid-Based Models

EPSANet [61] utilizes multiscale convolutional kernels in a pyramid structure and group convolution with different group sizes to earn richer multiscale feature representations and has achieved acceptable classification results on ImageNet. However, it cannot be effectively applied to hyperspectral classification because of the characteristics of HSIs. SSAN [21] extracts multiscale spectral information using variable convolution of $3 \times 3 \times q$ size, but cannot effectively consider the aggregation of multiscale spatial information. SMCN [23] and MCAL [33] innovate the combination of 3D multiscale feature acquisition and transformer. But ignores the attention distribution of multiscale feature information. AMFAN [22] and ELS2T [46] use three different dilate convolutions to generate multiscale features, where AMFAN gives different importance of multiscale features, but focuses only on spatial information and ignores spectral information. Our proposed PS2A module fully considers the extraction of spatial and spectral features and fuses the relevant attention information, which effectively solves the problems of the above models.

VI. Conclusion

In this study, a novel MHIAIFormer is proposed for the HSI classification. First, a PS2A feature extraction module is designed to efficiently extract multiscale information and enhance useful information in the image. Then, the extracted features from both spatial and spectral branches are adaptively fused by a SAWM. A novel embedding module named GMCD is proposed to learn feature information across latitudes and extract more discriminative spatial features without losing spectral information. In addition, we compensate for the multihead interaction capability of MHSA by utilizing additive attention and combining it with MHIA, and effectively extract global information relations in the spectral dimension. Finally, we use AMHCat to improve the integration mechanism of each attention head in the MHSA and generate the weights of the attention heads in different directions. Experimental results on four HSI datasets demonstrate that the MHIAIFormer effectively improves the classification accuracy and outperforms the state-of-the-art DL methods. In the future, we will work on building faster and lighter transformer structures for spectral-spatial HSI analysis. Moreover, DL methods such as unsupervised learning should be considered to reduce the need for HSI labeling.

References

[1] K. Tan, F. Wu, Q. Du, P. Du, and Y. Chen, "A parallel Gaussian–Bernoulli restricted Boltzmann machine for mining area classification with hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 2, pp. 627–636, Feb. 2019, doi: 10.1109/JSTARS.2019.2892975.

[2] T. C. W. Mok and A. C. S. Chung, "Fast symmetric diffeomorphic image registration with convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4643–4652, doi: 10.1109/CVPR42600.2020.00470.

[3] G. J. Lakshmi, V. Bodapati, S. S. Baji, and T. S. Panuganti, "Water quality monitoring using remote-sensing," in *Proc. Int. Conf. Comput. Intell. Sustain. Eng. Solutions*, 2023, pp. 624–628, doi: 10.1109/CISES58720.2023.10183395.

[4] R. Alamús et al., "Ground-based hyperspectral analysis of the urban nightscape," *ISPRS J. Photogram. Remote Sens.*, vol. 124, pp. 16–26, 2017, doi: 10.1016/j.isprsjprs.2016.12.004.

[5] Z. Xue, P. Du, J. Li, and H. Su, "Sparse graph regularization for robust crop mapping using hyperspectral remotely sensed imagery with very few in situ data," *ISPRS J. Photogram. Remote Sens.*, vol. 124, pp. 1–15, Feb. 2017, doi: 10.1016/j.isprsjprs.2016.12.003.

[6] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based *k*-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010, doi: 10.1109/TGRS.2010.2055876.

[7] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006, doi: 10.1109/TGRS.2006.877950.

[8] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005, doi: 10.1109/TGRS.2004.842481.

[9] M. Pal, "Multinomial logistic regression-based feature selection for hyperspectral data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 14, no. 1, pp. 214–220, 2012, doi: 10.1016/j.jag.2011.09.014.

[10] S. Jia, L. Shen, J. Zhu, and Q. Li, "A 3-D gabor phase-based coding and matching framework for hyperspectral imagery classification," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1176–1188, Apr. 2018, doi: 10.1109/TCYB.2017.2682846.

[11] M. Fauvel, J. Chanussot, J. A. Benediktsson, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2007, pp. 4834–4837, doi: 10.1109/IGARSS.2007.4423943.

[12] M. Pesaresi, A. Gerhardinger, and F. Kayitakire, "A robust built-up area presence index by anisotropic rotation-invariant textural measure," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 1, no. 3, pp. 180–192, Sep. 2008, doi: 10.1109/JSTARS.2008.2002869.

[13] X. Cao, L. Xu, D. Meng, Q. Zhao, and Z. Xu, "Integration of 3-dimensional discrete wavelet transform and Markov random field for hyperspectral image classification," *Neurocomputing*, vol. 226, pp. 90–100, Feb. 2017, doi: 10.1016/j.neucom.2016.11.034.

[14] A. Mughees and L. Tao, "Multiple deep-belief-network-based spectral-spatial classification of hyperspectral images," *Tsinghua Sci. Technol.*, vol. 24, no. 2, pp. 183–194, 2019, doi: 10.26599/TST.2018.9010043.

[15] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 120–147, Nov. 2018, doi: 10.1016/j.isprsjprs.2017.11.021.

[16] S. Jia, S. Jiang, S. Zhang, M. Xu, and X. Jia, "Graph-in-graph convolutional network for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 1157–1171, Jan. 2024, doi: 10.1109/TNNLS.2022.3182715.

[17] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017, doi: 10.1109/TGRS.2016.2636241.

[18] A. Ben Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018, doi: 10.1109/TGRS.2018.2818945.

[19] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018, doi: 10.1109/TGRS.2018.2794326.

[20] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020, doi: 10.1109/LGRS.2019.2918719.

[21] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020, doi: 10.1109/TGRS.2019.2951160.

[22] S. Zhang, J. Zhang, L. Xun, J. Wang, D. Zhang, and Z. Wu, "AM-FAN: Adaptive multiscale feature attention network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6012005, doi: 10.1109/LGRS.2022.3193488.

[23] H. Ge et al., "Pyramidal multiscale convolutional network with polarized self-attention for pixel-wise hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5504018, doi: 10.1109/TGRS.2023.3244805.

[24] Z. He, "Deep learning in image classification: A survey report," in *Proc. 2nd Int. Conf. Inf. Technol. Comput. Appl.*, 2020, pp. 174–177, doi: 10.1109/ITCA52113.2020.00043.

[25] G. Wu, W.-S. Zheng, Y. Lu, and Q. Tian, "PSLT: A light-weight vision transformer with ladder self-attention and progressive shift," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11120–11135, Sep. 2023, doi: 10.1109/TPAMI.2023.3265499.

[26] J.-C. Sheng, Y.-S. Liao, and C.-R. Huang, "Apply masked-attention mask transformer to instance segmentation in pathology images," in *Proc. 6th Int. Symp. Comput., Consum. Control*, 2023, pp. 342–345, doi: 10.1109/IS3C57901.2023.00098.

[27] X. Zhang, Y. Su, L. Gao, L. Bruzzone, X. Gu, and Q. Tian, "A lightweight transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5517617, doi: 10.1109/TGRS.2023.3297858.

[28] D. Yu, Q. Li, X. Wang, Z. Zhang, Y. Qian, and C. Xu, "DSTrans: Dual-stream transformer for hyperspectral image restoration," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3728–3738, doi: 10.1109/WACV56688.2023.00373.

[29] Y. Liu, J. Hu, X. Kang, J. Luo, and S. Fan, "Interactformer: Interactive transformer and CNN for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5531715, doi: 10.1109/TGRS.2022.3183468.

[30] F. Deng, W. Luo, Y. Ni, X. Wang, Y. Wang, and G. Zhang, "UMiT-Net: A U-shaped mix-transformer network for extracting precise roads using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5801513, doi: 10.1109/TGRS.2023.3281132.

[31] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5518615, doi: 10.1109/TGRS.2021.3130716.

[32] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5539014, doi: 10.1109/TGRS.2022.3207933.

[33] F. Xu, G. Zhang, C. Song, H. Wang, and S. Mei, "Multiscale and cross-level attention learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5501615, doi: 10.1109/TGRS.2023.3235819.

[34] H. Yu, Z. Xu, K. Zheng, D. Hong, H. Yang, and M. Song, "MSTNet: A multilevel spectral–spatial transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5532513, doi: 10.1109/TGRS.2022.3186400.

[35] D. Wang, J. Zhang, B. Du, L. Zhang, and D. Tao, "DCN-T: Dual context network with transformer for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 2536–2551, Apr. 023, doi: 10.1109/TIP.2023.3270104.

[36] J. Zou, W. He, and H. Zhang, "LESSFormer: Local-enhanced spectral-spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5535416, doi: 10.1109/TGRS.2022.3196771.

[37] J. Li, Z. Zhang, R. Song, Y. Li, and Q. Du, "SCFormer: Spectral coordinate transformer for cross-domain few-shot hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 33, pp. 840–855, Jan. 2024, doi: 10.1109/TIP.2024.3351443.

[38] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–Spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5522214, doi: 10.1109/TGRS.2022.3144158.

[39] R. Song, Y. Feng, W. Cheng, Z. Mu, and X. Wang, "BS2T: Bottleneck spatial–spectral transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5532117, doi: 10.1109/TGRS.2022.3185640.

[40] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral–spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5503615, doi: 10.1109/TGRS.2023.3242346.

[41] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5528715, doi: 10.1109/TGRS.2022.3171551.

[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[44] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13708–13717, doi: 10.1109/CVPR46437.2021.01350.

[45] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3138–3147, doi: 10.1109/WACV48630.2021.00318.

[46] S. Zhang, J. Zhang, X. Wang, J. Wang, and Z. Wu, "ELS2T: Efficient lightweight spectral–spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5518416, doi: 10.1109/TGRS.2023.3299442.

[47] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "SwiftFormer: Efficient additive attention for transformer-based real-time mobile vision applications," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17425–17436.

[48] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Oct. 2020, Accessed: Apr. 25, 2024. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[49] Y. Chen et al., "Mobile-former: Bridging MobileNet and transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5260–5269, doi: 10.1109/CVPR52688.2022.00520.

[50] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12165–12175, doi: 10.1109/CVPR52688.2022.01186.
[51] Z. Lu, H. Xie, C. Liu, and Y. Zhang, "Bridging the gap between vision transformers and convolutional neural networks on small datasets," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 14663–14677, 2022.
[52] W. Lin, Z. Wu, J. Chen, J. Huang, and L. Jin, "Scale-aware modulation meet transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 6015–6026.
[53] J. Bai et al., "Hyperspectral image classification based on multibranch attention transformer networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5535317, doi: 10.1109/TGRS.2022.3196661.
[54] E. Ouyang, B. Li, W. Hu, G. Zhang, L. Zhao, and J. Wu, "When multigranularity meets spatial–spectral attention: A hybrid transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 4401118, doi: 10.1109/TGRS.2023.3242978.
[55] J. Zhang, Z. Meng, F. Zhao, H. Liu, and Z. Chang, "Convolution transformer mixer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2022, Art. no. 6014205, doi: 10.1109/LGRS.2022.3208935.
[56] M. Ahmad, M. H. F. Butt, M. Mazzara, and S. Distifano, "Pyramid hierarchical transformer for hyperspectral image classification," 2024, *arXiv: 2404.14945*.
[57] Z. Zhao, D. Hu, H. Wang, and X. Yu, "Convolutional transformer network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Apr. 2022, Art. no. 6014205, doi: 10.1109/LGRS.2022.3169815.
[58] T. Arshad and J. Zhang, "Hierarchical attention transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, Mar. 2024, Art. no. 5504605, doi: 10.1109/LGRS.2024.3379509.
[59] X. Zhang et al., "Spectral–spatial self-attention networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5512115, doi: 10.1109/TGRS.2021.3102143.
[60] W. Yu, P. Zhou, S. Yan, and X. Wang, "InceptionNeXt: When inception meets ConvNeXt," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*, 2024, pp. 5672–5683.
[61] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, 2023, pp. 541–557, doi: 10.1007/978-3-031-26313-2_33.

Delong Kong received the B.E. degree in computer science and technology, in 2022, from the College of Computer Science and Technology, Qingdao University, Qingdao, China, where he is currently working toward the M.E. degree in software engineering.

His research interests mainly include computer vision, deep learning, and hyperspectral image classification.

Jiahua Zhang received the Ph.D. degree in cartography and remote sensing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences (CAS), Beijing, China, in 1998.

From 1999 to 2001, he held a Postdoctoral position with the National Institute for Environmental Studies, Tsukuba, Japan. Since 2002, he has been a Professor with the Chinese Academy of Meteorological Sciences, Beijing, China. Since 2012, he has been a Full Professor with the Institute of Remote Sensing and Digital Earth, CAS, and currently a Professor with Qingdao University. He has authored or coauthored more than 200 peer-reviewed papers, 30 international conferences papers, and six books. His research interests include remote sensing and geosciences, vegetation dynamics, land use and land-cover classification, image processing, deep learning, and coastal environment remote sensing.

Shichao Zhang received the M.E. degree in software engineering, in 2021, from the College of Computer Science and Technology, Qingdao University, Qingdao, China, where he is currently working toward the Ph.D. degree in software engineering.

His research interests mainly include computer vision, deep learning, and hyperspectral image classification.

Xiang Yu received the M.S. degree in computer technology, in 2020, from the College of Computer Science and Technology, Qingdao University, Qingdao, China, where he is currently working toward the Ph.D. degree in software engineering.

His current research interests include Big Data analysis, deep learning, remote sensing, computer vision, and hyperspectral image classification.

Foyez Ahmed Prodhan received the Ph.D. degree in cartography and geographic information system from the University of Chinese Academy of Sciences, Beijing, China, in 2021.

He is currently an Associate Professor with the Department of Agricultural Extension and Rural Development under the Faculty of Agriculture, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur City, Bangladesh. His research interests mainly include crop yield and drought monitoring using remote sensing Big Data, deep learning, and hyperspectral image classification.