HyperEAST: An Enhanced Attention-Based Spectral–Spatial Transformer With Self-Supervised Pretraining for Hyperspectral Image Classification

Jialin Tang, Nan Ma, Chen Jia, Rui Tian, and Yanhui Guo, *Member, IEEE*

***Abstract*—Hyperspectral images (HSIs) are essential in geoscientific applications, such as resource exploration, precision agriculture, and environmental monitoring, due to their rich spectral–spatial information. However, existing classification methods face notable limitations: Principal component analysis ignores spatial context, convolutional neural networks lack long-range modeling, and vision transformer (ViT)-based models often overfit under label-scarce conditions due to their high capacity and modality-agnostic design. To address these challenges, we propose HyperEAST, an efficient dual-branch ViT framework that explicitly decouples spectral and spatial feature modeling. At its core is a novel linear fusion attention mechanism, which replaces dot-product attention with a softmax-free additive formulation based on lightweight convolutions, enabling local–global representation learning with linear complexity. To enhance robustness under limited labels, we adopt a modality-aware masked image modeling strategy that separately reconstructs masked spectral and spatial tokens during self-supervised pretraining. We further introduce a dataset-aware hybrid loss combining cross-entropy and focal loss to mitigate class imbalance and sharpen decision boundaries. Experiments on four benchmark HSI datasets—WHU-Hi-HC, WHU-Hi-LK, Indian Pines, and Pavia University—demonstrate that HyperEAST achieves competitive accuracy, efficiency, and robustness.**

***Index Terms*—Hyperspectral image (HSI) classification, linear fusion attention, self-supervised learning (SSL), vision transformer (ViT).**

I. Introduction

HYPERSPECTRAL imaging (HSI) captures hundreds of contiguous spectral bands per pixel, enabling material-level classification across applications, such as environmental monitoring, precision agriculture, and mineral exploration [1]. However, its high dimensionality, spectral–spatial heterogeneity, and limited labeled data make accurate classification challenging [2], [3].

Early methods such as principal component analysis (PCA) and support vector machines (SVMs) relied on shallow modeling, while convolutional neural networks (CNNs) improved local pattern extraction [4], [5]. 2-D CNNs treat spectral bands as channels, whereas 3-D CNNs jointly model spectral–spatial information [6], [7]. Despite their success, CNNs are limited in capturing long-range dependencies due to localized receptive fields.

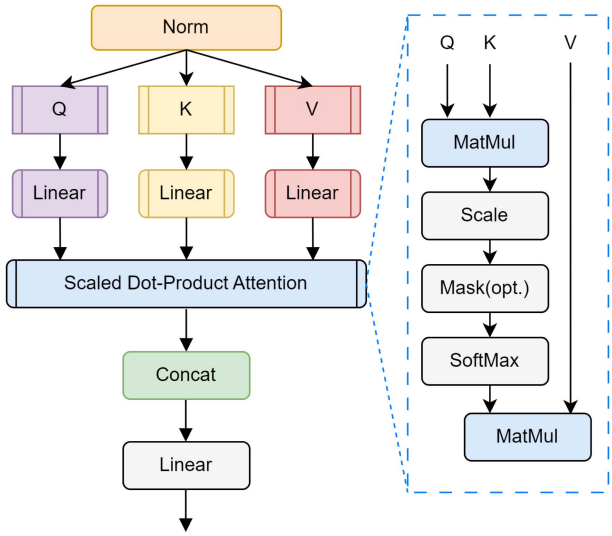

Fig. 1. Architecture of MHSA.

Transformers, with their global modeling capabilities, have been introduced to HSI classification. Their core operation—multihead self-attention (MHSA)—models token interactions via dot-product similarity followed by softmax normalization Fig. 1 [8]. While powerful, MHSA incurs quadratic complexity and lacks modality-specific inductive bias, making it inefficient for high-dimensional HSI data. Recent variants, such as MAEST and SpectralFormer, attempt to decouple modalities or enhance spectral awareness [9], [10], yet still face limitations in either efficiency or structural disentanglement.

To address these challenges, we propose **HyperEAST**, a dual-branch transformer architecture that explicitly separates spectral and spatial modeling. Central to our design is the linear fusion attention mechanism (LFAM), which replaces MHSA with a softmax-free, convolution-based additive operation. By leveraging grouped convolutions, LFAM introduces efficient, modality-aware interactions that better suit HSI data.

Received 10 June 2025; revised 10 July 2025 and 26 July 2025; accepted 14 August 2025. Date of publication 19 August 2025; date of current version 15 September 2025. This work was supported by the Shandong Provincial Natural Science Foundation under Grant ZR2023MF110. *(Corresponding author: Yanhui Guo.)*

The authors are with the School of Artificial Intelligence, Shandong Women's University, Jinan 250300, China (e-mail: jaylentang00@gmail.com; manan@sdwu.edu.cn; jiachen_1991@163.com; tianrr@foxmail.com; guoyanhui03@163.com).

Code is available at https://github.com/JaylenTang/HyperEAST

Digital Object Identifier 10.1109/JSTARS.2025.3599855

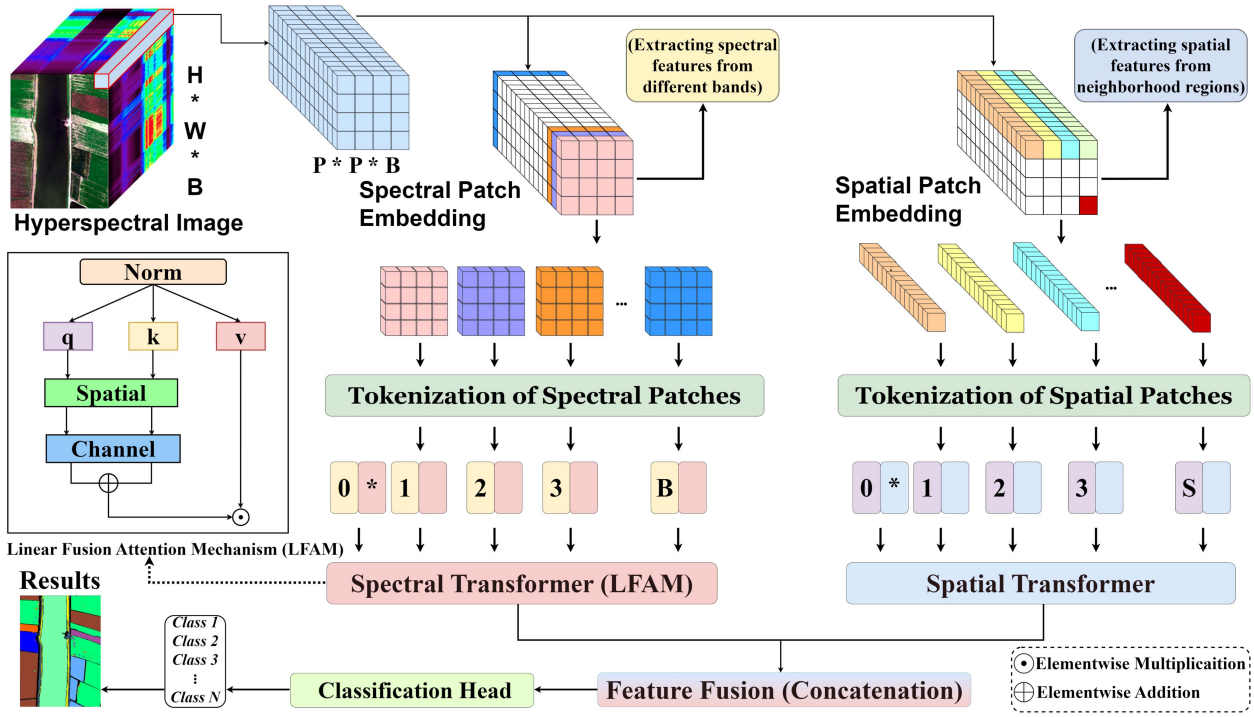

Fig. 2. Overview of the proposed HyperEAST architecture: A dual-branch transformer architecture with LFAM for HSIs. HyperEAST divides HSI data into spectral and spatial patches, tokenizes them separately, and processes them concurrently through a spatial transformer and a spectral transformer (equipped with LFAM), effectively capturing spectral and spatial features. The extracted features are concatenated and passed into an MLP classifier for final classification.

In addition, we incorporate a modality-aware self-supervised pretraining strategy based on masked image modeling (MIM), in which spectral and spatial tokens are masked and reconstructed independently. This promotes robust, disentangled representations under limited annotations [11], [12]. To mitigate class imbalance and improve decision boundary sensitivity, we design a hybrid loss (HLoss) combining cross-entropy (CE) and focal loss (FL) [13].

Our key contributions are as follows:

1) We propose LFAM, an efficient attention mechanism that eliminates dot-product and softmax operations via grouped convolutional fusion, reducing complexity while enhancing local–global modeling.
2) We develop a modality-aware MIM pretraining strategy that independently masks spectral and spatial tokens to improve generalization under low-label scenarios.
3) We adopt a dataset-aware HLoss combining CE and FL to address class imbalance and sharpen class boundaries.

Extensive experiments on four benchmark datasets—WHU-Hi-HC, WHU-Hi-LK, Indian Pines (IP), and Pavia University (PaviaU)—demonstrate that *HyperEAST* achieves strong classification performance with improved efficiency and generalization. The overall architecture is illustrated in Fig. 2.

II. Related Work

A. Deep Learning in HSI Classification and the Rise of Vision Transformers (ViT)

HSI classification has significantly advanced with the development of deep learning techniques that effectively capture both spectral and spatial information. Among the earliest approaches, CNNs demonstrated strong performance by leveraging local receptive fields and weight sharing [5]. Initial methods applied 1-D CNNs along the spectral axis to extract per-pixel spectral features, while 2-D CNNs focused on spatial textures within individual bands. To jointly model spectral–spatial information, 3-D CNNs were introduced, enabling the direct processing of hyperspectral cubes [14]. However, the high computational cost of 3-D operations limited their scalability.

To address this, hybrid architectures were proposed. Notably, HybridSN combines 3-D and 2-D convolutions to extract spectral–spatial features in a more efficient manner [7]. These CNN-based models achieve good performance, especially in scenarios with limited labeled data. Nevertheless, they are inherently constrained by localized receptive fields and a lack of global context modeling, which restrict their ability to capture long-range dependencies across spectral bands and spatial regions. To overcome these limitations, transformer architectures were introduced in the natural language processing domain as a mechanism for modeling global contextual relationships. The original transformer uses stacked encoder–decoder layers with MHSA and feedforward networks to model full pairwise token interactions [8]. While highly effective in sequential tasks, the transformer's quadratic complexity with respect to input length poses challenges in high-dimensional domains, such as hyperspectral imaging.

The ViT adapts this architecture to image classification by splitting an image into fixed-size patches, flattening them into token sequences, and applying standard transformer blocks [15]. ViTs inherit the global modeling capability of transformer,

making them promising for hyperspectral data. However, they also introduce new challenges. Directly feeding raw hyperspectral cubes into a ViT results in high computational cost and fails to preserve spectral continuity. As a workaround, preprocessing methods such as PCA or band sampling are often applied to reduce dimensionality—at the cost of potentially discarding informative spectral content [16].

Furthermore, standard ViTs lack modality awareness. They treat hyperspectral patches as flat tokens, without explicitly distinguishing between spectral and spatial variations [17]. This entanglement, combined with the requirement for large labeled datasets and high computational demands, limits their effectiveness in real-world HSI scenarios.

To address these challenges, several ViT-based models have been specifically tailored for hyperspectral classification. SpectralFormer introduces a groupwise spectral embedding module to model local spectral continuity prior to tokenization, improving spectral representation quality [10]. MAEST adopts a modality-aware dual-branch architecture that explicitly separates spectral and spatial processing, thereby reducing redundancy and enhancing modeling efficiency [9]. Additional refinements include ELViT [18], which reduces complexity while capturing both global and local dependencies.

B. Self-Supervised Learning (SSL) and MIM in HSI

The annotation of HSIs is often expensive and labor-intensive, resulting in a scarcity of labeled data across widely used datasets, such as IP and Houston2013. This constraint has motivated the adoption of SSL strategies to extract rich representations from abundant unlabeled data [19].

Apart from reconstruction-based SSL, other unsupervised approaches such as max–min distance embedding have also been explored to enhance feature separability in label-deficient environments [20]. These methods aim to improve representation learning by maximizing interclass margins in the latent space, offering a complementary direction to reconstruction-based techniques.

Among various SSL paradigms, MIM has demonstrated strong potential by enabling models to reconstruct masked input regions and learn context-aware, modality-agnostic features [21].

Recent advances in computer vision—such as MAE and BEiT—have inspired the adaptation of MIM to the hyperspectral domain. Several MIM-based approaches have emerged to address the challenges unique to HSIs. For example, SS-MAE adopts a unified ViT backbone to support cross-sensor modalities, including hyperspectral and multispectral data [22]. SSA-MIM proposes a joint spatial–spectral masking strategy to exploit intermodal correlations. SpectralMAE focuses solely on spectral continuity by masking and reconstructing spectral tokens [23], while S3Former introduces a split-modality masking scheme that separates the attention across modalities [24].

Despite promising results, most of these methods apply generic ViT architectures and rely on uniform or modality-agnostic masking strategies. Such designs often overlook the intrinsic structural heterogeneity of HSI data—namely, the distinct statistical and semantic properties of spectral versus spatial information. This leads to entangled feature representations and degrades transferability to downstream classification tasks, particularly under label scarcity or class imbalance.

To address these limitations, we propose a modality-aware MIM framework that explicitly decouples spectral and spatial branches during pretraining. By independently masking and reconstructing each modality, our design encourages intramodal specialization and better preserves modality-specific representations. Furthermore, we employ efficient attention mechanisms tailored to each branch, reducing computational complexity while enhancing pretraining efficiency and representational disentanglement. The overall pipeline is illustrated in Fig. 3.

C. Hybrid Architectures and Efficient Transformers for HSI

With the increasing adoption of ViTs in HSI classification, achieving a balance between modeling capacity and computational efficiency has become a key concern. Although ViTs offer strong global context modeling, their quadratic complexity with respect to input length renders them inefficient for high-resolution or real-time HSI applications. This has driven interest toward hybrid architectures and efficient attention mechanisms customized for spectral–spatial data.

Hybrid CNN–transformer frameworks aim to leverage the strengths of both components: CNNs effectively capture local textures and spatial continuity, while transformers excel at modeling long-range dependencies. For instance, Zhao et al. [25] introduced a convolution-enhanced transformer that injects local features into the attention mechanism to improve classification accuracy. Arya et al. [26] proposed a 3D-CNN-transformer (3DCmT) framework for joint spectral-spatial representation learning, and Ahmad et al. [27] developed the pyramid hierarchical spatial-spectral transformer (PHSS-Transformer), a dual-branch framework that captures modality-specific features across multiple scales. While these designs enhance feature expressiveness, they often suffer from elevated computational costs and structural complexity.

To improve efficiency, transformer variants have been explored in the vision domain. Swin transformer reduces computational burden via a shifted window strategy, BiFormer adopts bilevel routing attention to prioritize salient tokens, and EfficientViT employs cascaded grouped attention for memory-efficient processing [28]. MobileViT fuses convolutional priors with ViT modules to optimize inference on edge devices [29]. However, most of these models are tailored for RGB imagery and fail to account for the unique spectral characteristics of HSI, limiting their transferability.

Recent works have begun to address this gap by adapting efficient transformer architectures to the HSI domain. SwinMSP extends the Swin transformer with masked spectral pretraining to better capture spectral continuity [30]. MambaHSI, inspired by state-space modeling, introduces a linear-complexity spectral–spatial backbone [31]. FactoFormer further explores modality-aware design by adopting a dual-branch transformer that explicitly decouples spectral and spatial modeling, demonstrating the benefits of architectural disentanglement for HSI

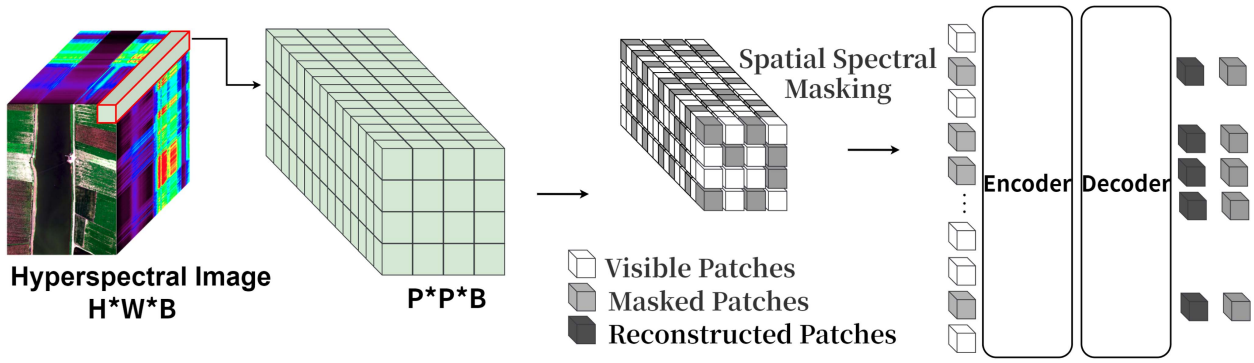

Fig. 3. Overview of the proposed masked spatial–spectral transformer, which randomly masks portions of spatial–spectral patches from HSIs and reconstructs these masked patches through encoder–decoder training.

classification [32]. Building on this trajectory, Yang et al. [33] proposed ACTN, an adaptive coupling transformer network that fuses a CNN and a transformer in parallel, allowing mutual interaction between local and global features. This design achieves more expressive representation and improved performance while maintaining efficiency. Similarly, Varahagiri et al. [34] proposed the convolution guided spectral-spatial transformer (3D-ConvSST), which integrates 3D-CNN with Transformer-based self-attention to enhance spectral–spatial fusion while reducing complexity. These approaches demonstrate promising directions for reconciling accuracy and computational demand in large-scale HSI classification. However, these models still rely on standard dot-product attention, lack efficient fusion mechanisms, or do not integrate lightweight inductive priors.

Motivated by these limitations, we propose *HyperEAST*, a dual-branch efficient transformer that explicitly separates spectral and spatial modeling. At its core is an **LFAM**, which replaces dot-product attention with grouped convolution and additive token mixing. This design enables localized, efficient interaction across modalities while maintaining strong discriminative capacity under limited supervision. By combining dual-path representation learning with efficient attention tailored to the structure of HSI data, HyperEAST achieves a favorable tradeoff between accuracy, generalization, and computational cost.

III. Methodology

A. Linear Fusion Attention Mechanism

1) Motivation: Traditional transformer architectures rely on dot-product self-attention ($\mathbf{QK}^\top$) to model token interactions. However, this operation introduces quadratic complexity $\mathcal{O}(N^2\,C)$, which is computationally expensive for HSIs due to their high spectral dimensionality and large spatial resolution. Moreover, the high capacity of MHSA may lead to overfitting under limited-label conditions, which is a common scenario in HSI classification.

To address these challenges, we propose the **LFAM**, an efficient alternative to MHSA. Inspired by recent developments in additive attention mechanisms, particularly CAS-ViT [35], LFAM replaces global dot-product interaction with a convolution-guided additive scheme. It leverages depthwise and pointwise convolutions to capture spatial and channel dependencies within local token neighborhoods, enabling efficient token mixing with linear complexity.

2) Design Overview: As shown in Fig. 4, LFAM consists of the following key stages.

1) *Query–key–value projection:* The input sequence is linearly projected into $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times C\times N}$ for further processing.
2) *Depthwise spatial convolution:* A depthwise 3×1 convolution is applied to $\mathbf{Q}+\mathbf{K}$ to enhance spatial–local context.
3) *Pointwise channel mixing:* A 1×1 convolution integrates channelwise information and outputs an attention map via sigmoid activation.
4) *Gated fusion:* The attention map modulates the value embedding through gated elementwise product, followed by projection and dropout.

3) Formulation: Given an input sequence $\mathbf{X} \in \mathbb{R}^{B\times N\times C}$, we first compute

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \text{Linear}(\mathbf{X}). \tag{1}$$

We transpose $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ into shape $\mathbb{R}^{B\times C\times N}$ and define the fused representation

$$\mathbf{F}_{\text{attn}} = \sigma\left(\mathbf{W}_2 * \text{ReLU}\left(\mathbf{W}_1 * (\mathbf{Q}+\mathbf{K})\right)\right) \tag{2}$$

where $*$ denotes convolution, $\mathbf{W}_1$ is a depthwise 3×1 kernel, and $\mathbf{W}_2$ is a 1×1 pointwise kernel. The attention-modulated output is then

$$\mathbf{O} = \tanh(\mathbf{F}_{\text{attn}}) \odot \mathbf{V}. \tag{3}$$

Finally, we apply a linear projection and dropout

$$\mathbf{O}_{\text{out}} = \text{Dropout}(\text{Linear}(\mathbf{O}^T)). \tag{4}$$

4) Computational Complexity: Compared with the quadratic cost of dot-product attention $\mathcal{O}(N^2\,C)$, LFAM reduces the complexity to

$$\mathcal{O}(\text{LFAM}) = \mathcal{O}(NC^2) + \mathcal{O}(kCN) \tag{5}$$

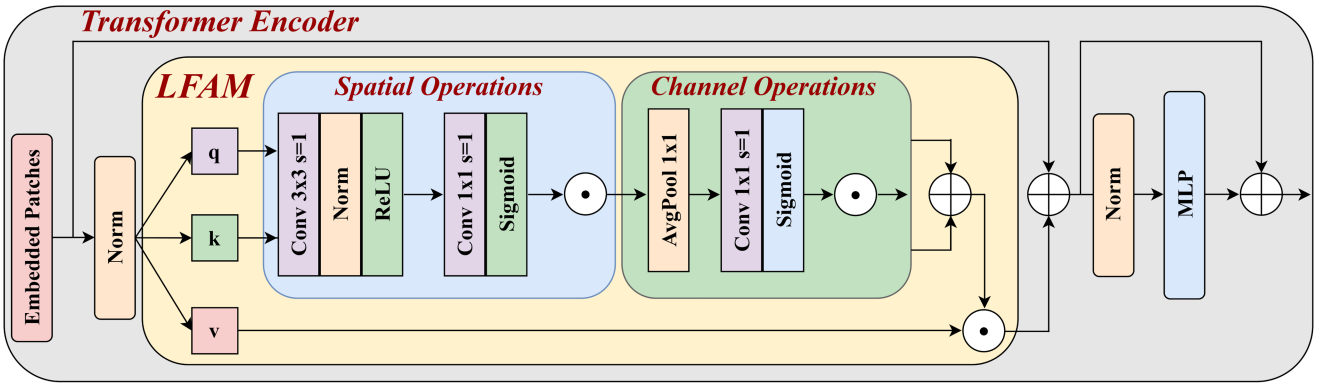

Fig. 4. Detailed architecture of the transformer encoder integrating the LFAM. Unlike conventional self-attention, LFAM linearly fuses Query ($\mathbf{q}$) and Key ($\mathbf{k}$) via parallel spatial and channel attention modules consisting of convolution, normalization, activation, pooling, and elementwise multiplications. The spatially and channel-enhanced features are then combined and modulate the value ($\mathbf{v}$) through elementwise multiplication, effectively capturing complex spatial–channel interactions in HSI patches.

where k is the convolution kernel size. This linear scaling greatly improves inference efficiency and scalability to large HSIs.

5) Advantages of LFAM:

1) *Linear complexity:* LFAM avoids pairwise attention computation and scales linearly with sequence length.
2) *Local dependency modeling:* Convolutional operations capture spatial continuity and contextual cues.
3) *Overfitting resistance:* The additive structure with gated modulation acts as an implicit regularizer.
4) *Hardware efficiency:* The design favors deployment on edge devices with high compatibility to ONNX/TensorRT.

B. Dual-Branch Spectral–Spatial Transformer With Modality-Aware MIM Pretraining

To fully exploit the complementary spectral and spatial characteristics of HSIs, we propose a dual-branch spectral–spatial transformer, as illustrated in Fig. 3. The architecture comprises two parallel and specialized branches—spectral and spatial—that independently extract domain-specific features prior to a final fusion stage for classification. We further introduce a modality-aware SSL strategy based on MIM, which enhances the model's representational power in low-label scenarios [36], [37].

1) Hybrid Attention Integration Strategy: A key innovation of our framework lies in its selective integration of attention mechanisms. Specifically, we apply the LFAM exclusively to the spectral branch, as efficient modeling of long-range spectral dependencies is critical in HSI analysis. Conversely, the spatial branch retains standard MHSA to preserve the ability to model fine-grained spatial structures. This hybrid design balances computational efficiency with representational power, enabling the model to perform robustly across diverse hyperspectral benchmarks.

2) Spectral Branch: The spectral branch captures interband correlations and long-range spectral dependencies. The input HSI cube $\mathbf{X} \in \mathbb{R}^{H\times W\times B}$ is divided into B nonoverlapping spectral patches along the spectral dimension. Each patch $\mathbf{a}_i \in \mathbb{R}^{H\times W\times 1}$ is flattened and projected into a fixed embedding dimension C_s

$$\mathbf{z}_{\text{spe},i} = \mathbf{E}_{\text{spe}} \cdot \text{Flatten}(\mathbf{a}_i) + \mathbf{e}_{\text{spe},i}, \quad i = 1, \ldots, B. \tag{6}$$

An additional `[CLS]` token $\mathbf{z}_{\text{cls,spe}}$ is prepended

$$\mathbf{Z}_{\text{spe}}^{(0)} = [\mathbf{z}_{\text{cls,spe}}, \mathbf{z}_{\text{spe},1}, \ldots, \mathbf{z}_{\text{spe},B}]. \tag{7}$$

This sequence is processed by transformer blocks with LFAM to enhance efficiency and spectral modeling.

3) Spatial Branch: The spatial branch focuses on local spatial patterns. The HSI cube is divided into $H \times W$ spatial patches, each $\mathbf{b}_i \in \mathbb{R}^{1\times 1\times B}$, which is flattened and projected

$$\mathbf{z}_{\text{spa},i} = \mathbf{E}_{\text{spa}} \cdot \text{Flatten}(\mathbf{b}_i) + \mathbf{e}_{\text{spa},i}, \quad i = 1, \ldots, H \cdot W. \tag{8}$$

With a `[CLS]` token prepended

$$\mathbf{Z}_{\text{spa}}^{(0)} = [\mathbf{z}_{\text{cls,spa}}, \mathbf{z}_{\text{spa},1}, \ldots, \mathbf{z}_{\text{spa},H\cdot W}]. \tag{9}$$

This is processed by standard transformer layers to extract spatial dependencies.

4) Spectral–Spatial Feature Fusion: Final layer tokens $\mathbf{z}_{\text{cls,spe}}^{(L)}$ and $\mathbf{z}_{\text{cls,spa}}^{(L)}$ are concatenated and passed to an MLP for classification

$$\mathbf{z}_{\text{fusion}} = \text{Concat}(\mathbf{z}_{\text{cls,spe}}^{(L)}, \mathbf{z}_{\text{cls,spa}}^{(L)}) \tag{10}$$

$$\hat{y} = \text{MLP}(\mathbf{z}_{\text{fusion}}). \tag{11}$$

Advantages:

1) *Targeted feature extraction:* Specialized processing reduces modality interference.
2) *Efficiency:* LFAM reduces spectral computation; spatial side preserves resolution.
3) *Complementary fusion:* Captures high-order spectral–spatial correlations.

5) Modality-Aware MIM: To address label scarcity, we adopt self-supervised pretraining via MIM. Each branch is pretrained independently using a factorized masking strategy.

- *Spectral branch:* Randomly mask subsets of spectral bands per pixel to capture spectral continuity.
- *Spatial branch:* Mask spatial patches to encourage learning of spatial dependencies.

Let $\mathcal{M}$ be the set of masked tokens. The reconstruction loss is defined as follows:

$$\mathcal{L}_{\text{MIM}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2 . \tag{12}$$

Each branch employs a dedicated reconstruction head to predict the masked tokens, promoting disentangled and specialized representations.

Benefits:

1) *Label-free learning:* Enables pretraining on large unlabeled HSI datasets.
2) *Branch-specific enhancement:* Ensures each branch learns complementary features.
3) *Better fine-tuning:* Leads to improved initialization and higher downstream accuracy.

C. HLoss Strategy for Class Imbalance Mitigation

1) Motivation: Class imbalance is a persistent challenge in HSI classification, where dominant land cover types often overwhelm minority classes. Recent studies have validated the effectiveness of FL and HLoss strategies in addressing class imbalance for HSI classification tasks [38]. Although both CE and FL are widely used individually, we observe that they exhibit complementary behavior:

CE provides stable learning for majority classes,

FL better emphasizes hard-to-classify and minority samples.

Motivated by this, we propose an HLoss strategy that linearly combines CE and FL, forming a more robust and generalizable optimization objective.

2) HLoss Formulation: Our hybrid objective combines the CE loss with the FL to enhance class-level discrimination. It is defined as

$$\mathcal{L}_{\text{hybrid}} = \lambda \cdot \mathcal{L}_{\text{CE}} + (1 - \lambda) \cdot \mathcal{L}_{\text{focal}}. \tag{13}$$

The FL component is formulated as

$$\mathcal{L}_{\text{focal}} = -\sum_{i=1}^{C} \alpha_i (1 - p_i)^{\gamma} \log(p_i) \tag{14}$$

where p_i is the predicted probability for class i, α_i denotes the inverse-frequency weight, and γ controls the emphasis on hard examples. While the CE term facilitates stable optimization, the focal term increases sensitivity to minority and difficult classes.

3) Dataset-Aware Weight Selection: Instead of fixing the weighting factor λ, we perform a grid search over various CE–FL ratios (e.g., 0.9–0.1, 0.6–0.4, 0.2–0.8) for each dataset. Experimental results reveal consistent trends as follows.

1) On relatively balanced datasets, CE-dominant combinations (e.g., $\lambda = 0.8$) offer better convergence and training stability.
2) On imbalanced datasets (e.g., IP), focal-dominant settings (e.g., $\lambda = 0.2$) significantly improve classification performance, especially for minority classes.
3) Across all datasets, the HLoss consistently outperforms standalone CE or FL, validating its robustness and practical effectiveness.

These findings suggest that a simple, dataset-aware weighted combination provides a flexible and effective solution for handling varying degrees of class imbalance.

4) Benefits:

1) *Complementary optimization:* Leverages CE's stability and FL's focus on hard samples.
2) *Dataset adaptability:* Optimal CE–FL ratio varies across datasets, reflecting intrinsic class distribution.
3) *Empirically validated:* Achieves higher OA and Kappa than single-loss baselines in all benchmark settings.

5) Conclusion: We present HyperEAST, a novel dual-branch ViT framework tailored for HSI classification. HyperEAST addresses key challenges in HSI analysis, including high spectral dimensionality, limited labeled data, and severe class imbalance. Recent work such as GSPST [39] enhances spectral modeling by incorporating group-spectral superposition and position-aware self-attention mechanisms. In contrast, our method achieves the same goal with a lightweight additive attention design (LFAM), reducing computation while preserving performance. The model explicitly decouples spectral and spatial representation learning through a two-branch architecture, enabling specialized modeling of long-range spectral dependencies and localized spatial patterns.

To enhance computational efficiency, we introduce the LFAM, a lightweight alternative to dot-product attention that employs additive convolutional transformations. Despite a slight increase in parameters and FLOPs, LFAM significantly reduces inference time and improves accuracy in the spectral branch, demonstrating an effective tradeoff between performance and complexity. This aligns with recent efforts such as SCluster-Former (Fang et al. [40], 2025), which leverage deformable convolutions and clustering priors to reduce redundancy in high-resolution HSI scenes.

To mitigate reliance on annotated data, we adopt a self-supervised MIM strategy for pretraining in spectral–spatial representation. During fine-tuning, we employ an HLoss that combines CE and FL to address class imbalance and improve robustness. Our self-supervised training framework builds on the idea that combining 3-D convolutions with ViT—seen in 3DVT (Su and Shao [41], 2025)—can improve feature fusion in low-label scenarios.

Extensive experiments on four benchmark datasets confirm that HyperEAST achieves competitive performance in terms of overall accuracy (OA), average accuracy (AA), and Kappa coefficient, while also reducing inference latency. In summary, HyperEAST offers a scalable, label-efficient, and deployment-ready solution for practical HSI classification, making it well-suited for real-world applications in remote sensing and geospatial data mining.

TABLE I
CLASS NAMES, COLOR LABELS, AND NUMBER OF SAMPLES FOR ALL FOUR DATASETS: IP, WHU-HI-HC, PAVIAU, AND WHU-HI-LK

IP			WHU-Hi-HC			Pavia U			WHU-Hi-LK		
No.	**Name**	**Total**	**No.**	**Name**	**Total**	**No.**	**Name**	**Total**	**No.**	**Name**	**Total**
C1	Alfalfa	46	C1	Strawberry	44 735	C1	Asphalt	6631	C1	Corn	34 511
C2	Corn-notill	1428	C2	Cowpea	22 753	C2	Meadows	18 649	C2	Cotton	8374
C3	Corn-mintill	830	C3	Soybean	10 287	C3	Gravel	2099	C3	Sesame	3031
C4	Corn	237	C4	Sorghum	5353	C4	Trees	3064	C4	Broad-leaf soybean	63 212
C5	Grass-pasture	483	C5	Water spinach	1200	C5	Painted metal sheets	1345	C5	Narrow-leaf soybean	4151
C6	Grass-trees	730	C6	Watermelon	4533	C6	Bare soil	5029	C6	Rice	11 854
C7	Grass-pasture-mowed	28	C7	Greens	5903	C7	Bitumen	1330	C7	Water	67 056
C8	Hay-windrowed	478	C8	Trees	17 978	C8	Self-blocking bricks	3682	C8	Roads and houses	7124
C9	Oats	20	C9	Grass	9469	C9	Shadows	947	C9	Mixed weed	5229
C10	Soybean-notill	972	C10	Red roof	10 516						
C11	Soybean-mintill	2455	C11	Gray roof	16 911						
C12	Soybean-clean	593	C12	Plastic	3679						
C13	Wheat	205	C13	Bare soil	9116						
C14	Woods	1265	C14	Road	18 560						
C15	Buildings-G-T-drives	386	C15	Bright object	1136						
C16	Stone-steel-towers	93	C16	Water	75 401						
Total		**10 249**	**Total**		**257 530**	**Total**		**42 766**	**Total**		**204 542**

IV. EXPERIMENTS

A. Hyperspectral Datasets

To comprehensively evaluate the effectiveness and generalization capability of our proposed method, we conduct experiments on four widely used hyperspectral datasets: IP, WHU-Hi-HanChuan (WHU-Hi-HC), WHU-Hi-LongKou (WHU-Hi-LK), and PaviaU. These datasets encompass diverse spectral resolutions, spatial scales, and scene complexities, making them well-suited benchmarks for HSI classification.

IP and PaviaU are classic benchmarks in agricultural and urban settings, respectively, and have been extensively used in prior works [42], [43]. WHU-Hi-HC and WHU-Hi-LK, on the other hand, provide large-scale UAV-acquired imagery with heterogeneous land cover, making them ideal for evaluating spectral–spatial learning models [44]. We follow the same train-test splits as adopted in SpectralFormer [10] and MAEST [9] to ensure fair comparisons.

1) Indian Pines: The IP dataset is a widely used benchmark in hyperspectral remote sensing. It was collected by the airborne visible/infrared imaging spectrometer over northwestern Indiana, USA, primarily covering agricultural land. The original data contains 220 spectral bands across wavelengths from 400 to 2500 nm; however, 20 bands affected by atmospheric absorption are typically removed, resulting in 200 valid bands. With a spatial dimension of 145×145 pixels and 16 land-cover classes (mainly crop types and forested areas), the dataset poses a classification challenge due to the high spectral similarity between certain crop species. Classwise sample statistics are summarized in Table I.

2) WHU-Hi-HanChuan: The WHU-Hi-HC dataset was acquired using a UAV-mounted hyperspectral imaging system over HanChuan, Hubei Province, China. It has a spatial resolution of 1217×303 pixels, with more than 283 000 labeled samples retained after preprocessing. The region includes diverse land-cover types, such as vegetation, water bodies, and urban settlements, making the dataset challenging due to interclass spectral similarity. Compared to WHU-Hi-LK, WHU-Hi-HC spans a larger area with greater heterogeneity, making it a suitable benchmark for evaluating spectral–spatial feature extraction. Class distribution details are listed in Table I.

3) Pavia University: The PaviaU dataset was captured by the reflective optics system imaging spectrometer over the urban area of Pavia, Italy. After removing noisy bands, 103 spectral channels are retained. The image has a spatial resolution of 610×610 pixels and includes nine land-cover classes, primarily consisting of urban structures such as roads, buildings, and vegetation. Unlike IP, which focuses on agricultural landscapes, PaviaU emphasizes manmade environments, making it a popular choice for evaluating spectral–spatial classification methods. Per-class sample counts are provided in Table I.

4) WHU-Hi-LongKou: The WHU-Hi-LK dataset is part of the WHU-Hi airborne hyperspectral series, collected using a Headwall Nano-Hyperspec sensor mounted on a DJI Matrice 600 Pro UAV. The data were acquired over Longkou Town, Hubei Province, China, on 17 July 2018, between 13:49 and 14:37 local time. It has a spatial size of 550×400 pixels, and 204 542 labeled samples are retained after background removal. The scene primarily comprises agricultural and built-up areas, serving as a valuable benchmark for assessing model performance on heterogeneous land-cover classification. Classwise sample statistics are listed in Table I.

B. Pretraining Strategy

To improve the model's ability to capture informative spectral–spatial representations, we adopt a self-supervised pretraining strategy based on MIM. This approach enables the model to learn rich features from unlabeled hyperspectral data by reconstructing randomly masked portions of the input, leveraging the inherent spectral correlations across bands.

Unlike fully supervised methods that rely heavily on large-scale labeled datasets, our self-supervised framework encourages the model to recover masked spectral tokens, promoting the learning of high-order spectral dependencies and improving generalization across diverse datasets.

We conduct pretraining on the unlabeled portion of each dataset, as summarized in Table II. Pretraining is conducted

TABLE II
DETAILS OF DATA USED FOR PRETRAINING, FINE-TUNING, AND TESTING ON FOUR HYPERSPECTRAL DATASETS

Dataset	Pretrain	Fine-tune	Test
IP	10 659	695	9671
Pavia U	163 477	3921	40 002
WHU-Hi-LK	15 458	900	203 642
WHU-Hi-HC	111 221	1600	255 930

for 200 epochs using the Adam optimizer with a batch size of 32 and a learning rate of 5×10^{-4}. We adopt a StepLR scheduler without weight decay. During pretraining, we apply a masking ratio of 70% to both the spectral (SPE) and spatial (SPA) branches. These hyperparameters are applied consistently across Swin-MSP, FactoFormer, and our proposed HyperEAST model to ensure fair comparison. For all other baselines, we follow their original training protocols as reported in their respective publications.

C. Fine-Tuning Setup

Following self-supervised pretraining, the model is fine-tuned on labeled hyperspectral datasets. During this stage, we incorporate several architectural and optimization enhancements to further improve classification performance, including the LFAM, an HLoss function combining CE and FL, and a dual-branch architecture that decouples spectral and spatial feature learning.

Fine-tuning is performed using the Adam optimizer with a StepLR learning rate scheduler, a batch size of 32, and no weight decay. Learning rates and training epochs are dataset-specific: for IP, we use 3×10^{-4} for 80 epochs; for PaviaU, 1×10^{-3} for 80 epochs; for Houston 2013, 2×10^{-3} for 40 epochs; and for WHU-Hi datasets, 1×10^{-3} for 40 epochs.

The spectral and spatial branches are configured with five transformer layers each. The spectral branch uses an embedding size of 32 and 3-D patches of size $7 \times 7 \times 1$, along with a lightweight MLP module with expansion factor 4, and replaces standard MHSA with the proposed LFAM to reduce computational overhead while preserving spectral representation capability. In contrast, the spatial branch uses an embedding size of 64, operates on $1 \times 1 \times B$ 2-D patches, and is equipped with four attention heads and an MLP expansion factor of 8.

The details of the data splits used for pretraining, fine-tuning, and testing are listed in Table II. Evaluation is conducted on four benchmark hyperspectral datasets: WHU-Hi-HC, WHU-Hi-LK, IP, and PaviaU. To ensure fair and reproducible comparisons, we follow a consistent train–test split protocol across all experiments.

All experiments were conducted on a workstation equipped with a 24 GB NVIDIA RTX 3090 GPU. In addition, we utilized GPU warmup and CUDA synchronization to maintain the consistency of the results.

D. HLoss Optimization and LFAM Ablation

a) Computational Efficiency: In addition to accuracy improvements, HyperEAST achieves notable computational efficiency, reducing both training and inference time compared to conventional transformer-based baselines. As shown in Table III, LFAM improves classification performance while introducing manageable overhead, especially when applied selectively to the spectral branch. This tradeoff is particularly beneficial for real-time or resource-constrained remote sensing applications.

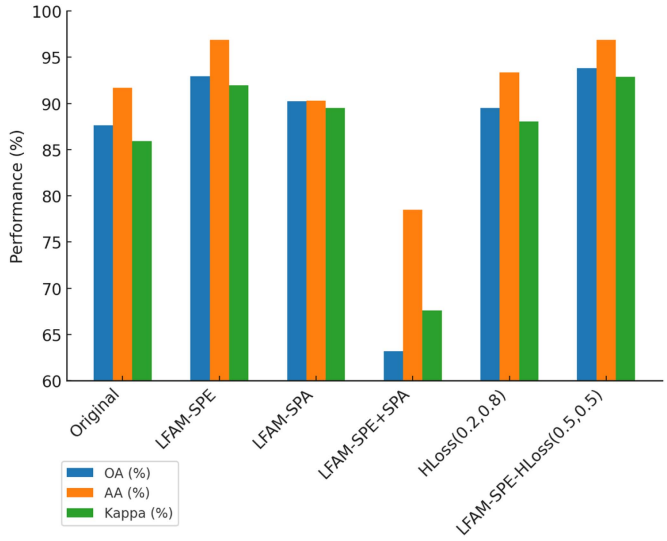

Fig. 5. Performance comparison of OA, AA, and Kappa coefficient for different configurations of LFAM and HLoss on the IP dataset.

To further investigate the contribution of LFAM and the HLoss design, we conduct a series of ablation studies. First, we evaluate five configurations for LFAM usage:

1) applied to the spectral branch only (LFAM-SPE);
2) to the spatial branch only (LFAM-SPA);
3) to both branches (LFAM-SPE+SPA);
4) HLoss without LFAM;
5) LFAM-SPE combined with the optimal HLoss at a 0.5:0.5 ratio.

Results are presented in Table IV and illustrated in Fig. 5 . The findings confirm that LFAM significantly boosts spectral modeling, with LFAM-SPE alone improving OA from 87.66% to 92.95% .

Meanwhile, to mitigate class imbalance and enhance robustness, we employ an HLoss function that linearly combines CE and FL

$$L_{\text{total}} = \alpha \mathcal{L}_{\text{FL}} + \beta \mathcal{L}_{\text{CE}} \tag{15}$$

where $\alpha + \beta = 1$, and the weighting is selected according to the imbalance characteristics of each dataset. On the IP dataset, shown in Table V, the best performance is achieved with $\alpha = 0.2$ and $\beta = 0.8$, resulting in 89.52% OA, 93.37% AA, and a Kappa of 0.8804. These results, along with visualizations in Fig. 6, demonstrate that even a modest incorporation of FL significantly improves performance on minority classes.

When combining LFAM-SPE with the optimal HLoss, HyperEAST achieves its best performance: 93.81% OA, 96.89% AA, and 0.9291 Kappa. This balance between LFAM's representational power and the HLoss's imbalance resilience validates our overall design strategy.

TABLE III
CLASSIFICATION PERFORMANCE AND COMPUTATIONAL EFFICIENCY ON WHU-HI-HC AND WHU-HI-LK DATASETS

Method	WHU-Hi-HC							WHU-Hi-LK						
	OA(%)	AA(%)	Kappa	Time	GPU Memory	FLOPs	Params	OA	AA	Kappa	Time	GPU Memory	FLOPs	Params
BASELINE	92.17	91.07	0.9087	48.62 s	564.81 MB	12.11 M	0.16 M	98.19	98.71	0.9763	38.49 s	554.60 MB	12.00 M	0.15 M
LFAM-SPE	91.00	90.50	0.8952	42.25 s	356.82 MB	41.76 M	0.55 M	98.51	97.69	0.9805	33.14 s	362.30 MB	41.46 M	0.53 M
LFAM-SPA	82.28	80.79	0.7949	51.99 s	705.64 MB	41.18 M	0.54 M	93.54	93.31	0.9165	39.57 s	695.55 MB	40.90 M	0.54 M
LFAM-SPE+SPA	76.03	71.44	0.7230	43.63 s	359.34 MB	42.90 M	0.55 M	86.87	83.57	0.8330	33.91 s	359.16 MB	42.59 M	0.55 M
HLoss	92.89	90.67	0.9169	48.15 s	564.81 MB	12.11 M	0.16 M	98.51	98.72	0.9804	37.80 s	554.60 MB	12.00 M	0.15 M
LFAM-SPE+HLoss	**93.00**	**91.82**	**0.9183**	**40.96 s**	**356.82 MB**	**41.76 M**	**0.55 M**	**98.87**	**98.75**	**0.9851**	**32.25 s**	**362.30 MB**	**41.46 M**	**0.53 M**

Each experiment is repeated ten times with different random seeds. Standard deviations are negligible (<1%) and thus omitted for brevity.
Bold values indicate the best performance among compared methods.

TABLE IV
CLASSIFICATION PERFORMANCE (OA, AA, AND KAPPA) AFTER FINE-TUNING DIFFERENT CONFIGURATIONS OF LFAM AND HLOSS ON THE IP DATASET

Dataset (IP)	OA (%)	AA (%)	Kappa
Original	87.66	91.72	0.8593
LFAM-SPE	92.95	96.87	0.9195
LFAM-SPA	90.25	90.28	0.8954
LFAM-SPE+SPA	63.18	78.52	0.6759
HLoss(0.2,0.8)	89.52	93.37	0.8804
LFAM-SPE+HLoss(0.5,0.5)	**93.81**	**96.89**	**0.9291**

Bold values indicate the best performance among compared methods.

TABLE V
CLASSIFICATION PERFORMANCE (OA, AA, AND KAPPA) OBTAINED USING DIFFERENT CE AND FL RATIO COMBINATIONS (HLOSS) ON THE IP DATASET

CE-FL	OA	AA	Kappa	CE-FL	OA	AA	Kappa
Original	0.8766	0.9172	0.8593	0.5,0.5	0.8630	0.9237	0.8448
0.9,0.1	0.8940	0.9248	0.8790	0.4,0.6	0.8943	**0.9381**	0.8798
0.8,0.2	0.8808	0.9302	0.8643	0.3,0.7	0.8840	0.9360	0.8679
0.7,0.3	0.8576	0.9138	0.8381	**0.2,0.8**	**0.8952**	0.9337	**0.8804**
0.6,0.4	0.8836	0.9221	0.8675	0.1,0.9	0.8888	0.9301	0.8732

Bold values indicate the best performance among compared methods.

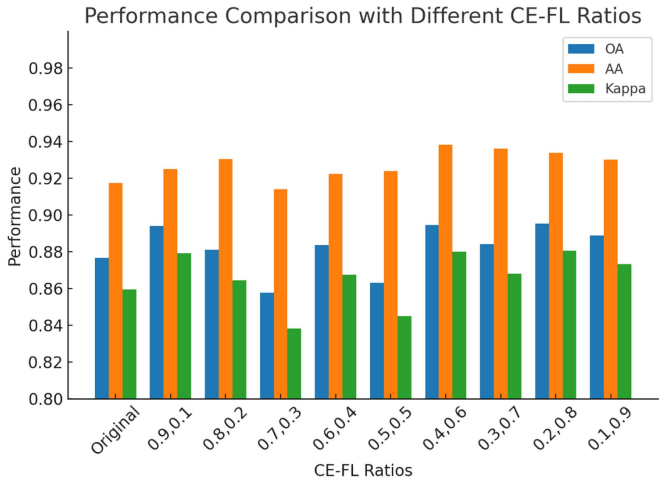

Fig. 6. Performance comparison of OA, AA, and Kappa coefficient under different CE and FL ratios on the IP dataset.

E. Component Effectiveness and Comparative Evaluation

To validate the effectiveness of the proposed *HyperEAST* framework, we conduct comprehensive ablation studies, computational efficiency analysis, and comparative evaluations against state-of-the-art (SOTA) methods across four benchmark hyperspectral datasets. All experiments are conducted under a fixed data split and deterministic training settings to ensure fair and reproducible comparison. Each experiment is repeated ten times with different random seeds, and the average results are reported. Since the standard deviations are consistently negligible (typically less than 1%), we omit them from the tables for brevity.

a) Ablation Study and LFAM design justification: As shown in Table IV, removing either the HLoss or the LFAM results in a substantial decline in classification performance, underscoring their complementary contributions. The HLoss improves robustness to class imbalance, while LFAM enhances the efficiency and expressiveness of spectral feature modeling.

LFAM is specifically applied to the spectral transformer branch, given the high spectral dimensionality inherent in hyperspectral data. As spectral information plays a more critical role than spatial cues in most HSI tasks, restricting LFAM to the spectral branch strikes the optimal balance between performance and computational cost. Empirical results presented in Table IV further corroborates this design choice. Specifically, the LFAM-SPE variant—where LFAM is applied solely to the spectral branch—outperforms both LFAM-SPA (applied to the spatial branch) and LFAM-SPE+SPA (LFAM applied to both branches), achieving superior accuracy while maintaining lower computational cost. These findings reinforce our hypothesis that LFAM is particularly effective for spectral feature fusion due to the significant redundancy inherent among spectral bands. In contrast, spatial features are typically lower dimensional and less redundant, making conventional attention mechanisms adequate for their processing. Consequently, introducing LFAM into the spatial branch (LFAM-SPA) not only increases model complexity but also provides diminishing returns in terms of performance improvement.

b) Comparison with current competitive method: We compare HyperEAST with several representative hyperspectral classification models, including 2-D CNN [6], 3-D CNN [7], RCNN [45], GAHT [46], SpectralFormer [10], MAEST [9],

TABLE VI
CLASSIFICATION ACCURACY (%) COMPARISON OF THE PROPOSED HYPEREAST MODEL WITH SOTA METHODS ON THE IP DATASET

Class	2-D CNN	RCNN	3-D CNN	GAHT	SF (pixel)	SF (patch)	MAEST (pixel)	MAEST (patch)	Swin MSP	Facto former	(Ours)
1	**100**	**100**	97.78	**100**	74.19	90.32	96.77	93.55	73.23	83.96	87.5
2	96.68	96	96.09	76.57	54.21	39.19	80.48	66.04	65.71	96.05	**98.21**
3	52.84	54.04	52.65	75.16	38.08	73.21	70.38	84.62	76.74	94.02	**100**
4	46.45	45.45	52.17	93.23	76.47	93.05	86.10	**99.47**	98.13	98.93	99.11
5	66.30	68.12	71.77	91.90	75.52	69.75	88.68	89.84	90.83	**100**	**100**
6	99.86	**100**	**100**	97.61	88.82	88.38	94.85	90.15	95.78	99.09	99.54
7	**100**	**100**	96.30	**100**	84.62	92.31	76.92	**100**	99.23	98.46	93.36
8	70.81	70.81	70.81	**99.87**	92.06	89.25	86.45	95.56	99.28	98.83	90.57
9	**100**	**100**	**100**	**100**	60.00	**100**	80.00	**100**	**100**	**100**	90.25
10	91.15	90.42	90.09	71.70	86.33	78.96	80.37	86.77	77.80	99.38	**100**
11	92.94	92.80	92.48	66.39	61.87	35.76	51.64	60.67	58.74	94.69	**96.22**
12	75.73	76.51	73.68	81.06	39.78	70.90	53.22	78.45	76.30	**99.09**	95.45
13	**100**	**100**	**100**	99.68	98.71	**100**	96.13	99.35	99.81	99.74	**100**
14	80.57	81.13	80.60	91.02	87.08	78.85	90.29	85.10	92.38	93.84	**100**
15	37.34	36.82	36.44	91.80	46.13	66.37	49.70	62.80	82.62	81.25	**100**
16	95.56	97.83	92.13	**100**	**100**	**100**	97.67	**100**	**100**	**100**	**100**
OA (%)	73.89	73.89	73.83	80.47	67.59	62.46	73.15	76.84	76.82	87.66	**93.81**
AA (%)	81.64	81.87	81.44	89.75	72.74	79.14	79.98	87.02	86.66	91.72	**96.89**
Kappa (%)	71.03	71.04	70.96	77.86	63.00	58.50	69.70	73.87	73.87	85.93	**92.91**

Each experiment is repeated ten times with different random seeds. Standard deviations are negligible (<1%) and thus omitted for brevity.
Bold values indicate the best performance among compared methods.

TABLE VII
CLASSIFICATION ACCURACY (%) COMPARISON OF THE PROPOSED HYPEREAST MODEL WITH SOTA METHODS ON THE PAVIAU DATASET

Class	2-D CNN	RCNN	3-D CNN	GAHT	SF (pixel)	SF (patch)	MAEST (pixel)	MAEST (patch)	Swin MSP	Facto former	(Ours)
1	92.62	**93.43**	93.04	86.01	55.22	58.02	57.27	70.46	90.67	93.39	92.45
2	85.48	85.72	85.17	89.68	33.50	59.39	52.83	63.43	89.83	**97.83**	96.58
3	83.33	84.42	83.14	73.89	19.42	26.50	50.66	**93.27**	88.94	72.29	80.88
4	95.78	95.36	95.54	96.58	97.84	96.62	97.61	**98.24**	**98.24**	96.46	98.08
5	99.93	**100.00**	**100.00**	99.86	99.38	99.85	99.69	**100.00**	99.99	99.64	99.28
6	**100.00**	**100.00**	**100.00**	88.07	88.55	71.36	72.58	65.07	90.12	90.16	93.42
7	**100.00**	**100.00**	**100.00**	93.39	92.27	95.63	95.31	96.02	96.80	99.59	98.27
8	99.76	**99.84**	99.75	83.12	88.16	90.42	74.01	71.86	83.09	99.29	99.35
9	98.28	98.72	98.61	99.77	99.67	99.89	99.89	**100.00**	99.97	99.98	98.87
OA (%)	84.72	85.16	84.62	88.73	57.14	67.50	64.46	72.24	90.71	95.15	**95.36**
AA (%)	95.02	**95.28**	95.03	90.04	74.89	77.52	77.76	84.26	93.07	93.72	95.24
Kappa (%)	81.08	81.59	80.96	85.32	49.68	59.75	56.63	65.27	87.88	93.46	**93.76**

Each experiment is repeated ten times with different random seeds. Standard deviations are negligible (<1%) and thus omitted for brevity.
Bold values indicate the best performance among compared methods.

TABLE VIII
CLASSIFICATION ACCURACY (%) COMPARISON OF THE PROPOSED HYPEREAST MODEL WITH SOTA METHODS ON THE WHU-HI-HC DATASET

Class	2-D CNN	RCNN	3-D CNN	GAHT	SF (pixel)	SF (patch)	MAEST (pixel)	MAEST (patch)	Swin MSP	Facto former	(Ours)
1	93.05	93.00	**93.21**	89.29	58.48	76.00	74.20	86.84	80.72	90.49	89.62
2	94.20	94.48	94.28	82.67	47.98	43.12	34.07	41.77	70.17	95.14	**96.27**
3	94.11	94.32	94.20	82.87	57.47	87.52	18.34	34.52	80.58	88.91	**96.27**
4	94.27	94.52	94.48	96.77	96.95	95.79	48.93	80.46	95.48	98.59	**100.00**
5	97.61	97.87	98.65	98.50	63.22	97.52	34.17	72.52	96.04	99.55	**100.00**
6	91.35	92.02	91.79	72.33	20.39	40.20	15.90	34.26	56.19	91.88	**92.02**
7	**99.76**	99.69	99.75	87.81	38.18	37.95	20.78	37.31	87.44	89.35	93.35
8	**97.39**	**97.39**	97.29	96.64	84.49	96.16	36.16	95.65	61.24	94.03	95.34
9	98.49	**98.64**	98.49	88.47	49.41	91.38	96.49	90.11	68.54	94.03	95.34
10	84.13	85.23	85.26	75.41	41.05	68.09	38.74	55.95	**91.02**	81.18	87.27
11	89.44	89.20	89.02	87.39	67.96	80.54	65.23	76.82	87.58	92.17	**93.00**
12	**93.92**	92.87	93.13	85.80	64.24	78.20	58.88	70.05	76.46	91.07	91.82
13	94.55	**94.98**	94.67	85.31	63.54	77.51	60.42	73.15	66.46	90.87	91.83
14	90.61	90.97	90.50	87.39	76.20	72.56	73.68	34.20	71.66	94.78	**96.76**
15	96.73	96.07	96.73	97.90	80.48	87.23	84.44	88.49	93.78	92.78	**98.31**
16	93.21	93.26	93.12	97.61	82.24	96.23	75.49	93.68	97.44	**98.78**	98.71
OA (%)	87.44	87.63	87.48	87.39	67.96	80.54	65.23	76.82	82.69	92.17	**93.00**
AA (%)	93.93	94.03	**94.04**	85.80	64.24	78.20	58.88	70.05	80.05	91.07	91.82
Kappa (%)	85.61	85.83	85.66	85.31	63.54	77.51	60.42	73.15	79.92	90.87	**91.83**

Each experiment is repeated ten times with different random seeds. Standard deviations are negligible (<1%) and thus omitted for brevity.
Bold values indicate the best performance among compared methods.

Swin-MSP [30], and FactoFormer [32]. All baselines are implemented using their official codebases or faithfully reimplemented based on the settings reported in their original publications. Experiments are conducted on four standard datasets to ensure a fair and comprehensive comparison.

c) Overall performance evaluation: As summarized in Tables VI–IX, HyperEAST consistently outperforms all competing methods across all datasets in terms of OA, AA, and Kappa coefficient. It achieves the highest accuracy among all baselines and establishes a new competitive method in HSI

TABLE IX
CLASSIFICATION ACCURACY (%) COMPARISON OF THE PROPOSED HYPEREAST MODEL WITH SOTA METHODS ON THE WHU-HI-LK DATASET

Class	2-D CNN	RCNN	3-D CNN	GAHT	SF (pixel)	SF (patch)	MAEST (pixel)	MAEST (patch)	Swin MSP	Facto former	(Ours)
1	88.12	88.54	88.14	99.39	84.68	73.64	73.47	96.33	96.46	**99.86**	99.33
2	90.95	90.85	90.35	97.03	42.61	74.36	73.63	66.25	98.92	**99.85**	98.30
3	89.99	91.4	90.90	98.96	87.76	93.53	62.33	94.73	98.21	**99.65**	**99.65**
4	90.99	90.87	90.84	95.63	33.66	54.99	61.08	81.58	**98.15**	94.96	97.63
5	94.40	93.67	94.30	98.61	74.45	45.48	27.72	79.52	92.05	**99.75**	**99.75**
6	96.61	96.78	96.95	99.20	89.87	91.06	94.35	99.65	98.67	**99.99**	99.97
7	97.37	97.32	97.36	99.39	92.31	95.63	95.19	96.02	98.37	**99.98**	99.97
8	97.28	97.03	97.10	96.61	80.89	75.99	75.19	88.52	**98.99**	97.96	96.57
9	96.58	96.60	96.65	97.31	34.20	53.39	50.76	68.43	92.69	98.36	**98.46**
OA (%)	87.74	87.78	87.68	98.05	70.93	76.84	77.95	90.55	96.33	98.19	**98.87**
AA (%)	93.59	93.67	96.33	98.05	69.79	73.63	68.72	86.61	96.93	**99.75**	**99.75**
Kappa (%)	84.47	84.51	84.39	97.46	64.59	71.24	72.30	87.80	95.39	97.63	**98.51**

Each experiment is repeated ten times with different random seeds. Standard deviations are negligible (<1%) and thus omitted for brevity.
Bold values indicate the best performance among compared methods.

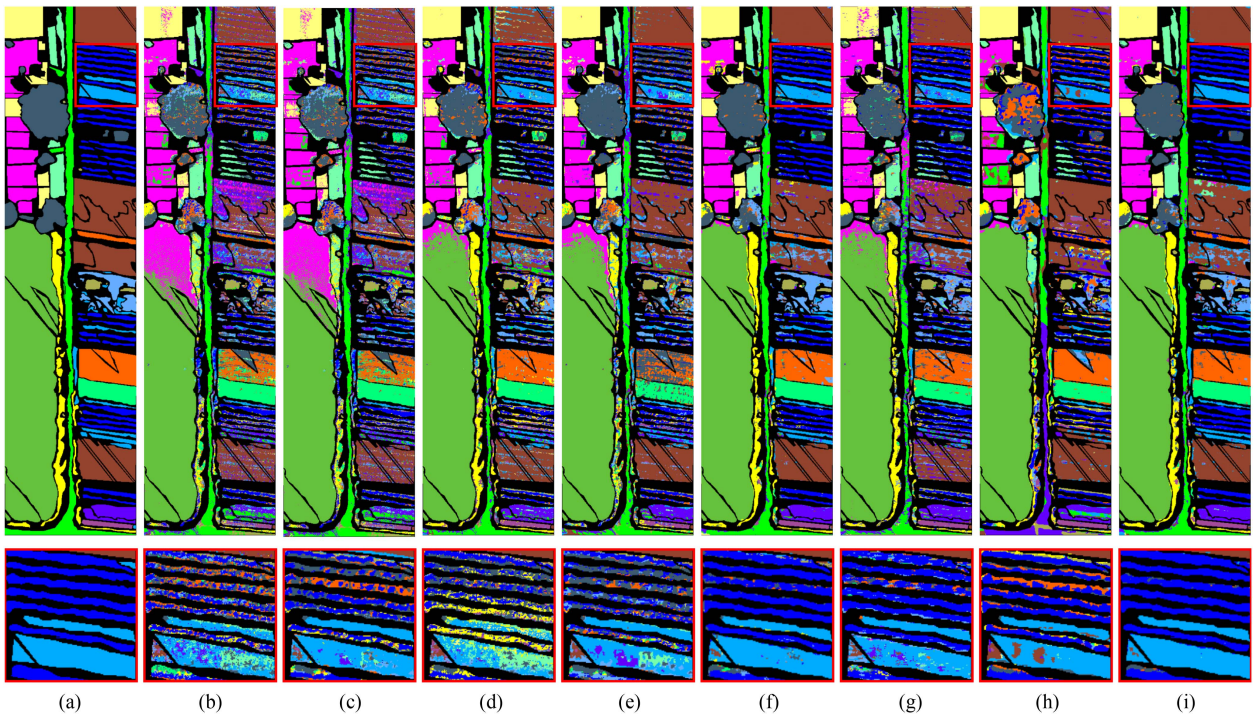

Fig. 7. Visualization of ground-truth and classification maps obtained by different models on the WHU-Hi-HC dataset (from left to right: (a) ground truth, (b) SF (pixel), (c) SF (patch), (d) MAEST (pixel), (e) MAEST (patch), (f) GAHT, (g) Swin-MSP, (h) FactoFormer, and (i) HyperEAST).

classification. Note that we only report the inference time in the main results table, as it directly reflects the model's practical efficiency. While the inclusion of LFAM modules introduces a moderate increase in parameter size (from 0.16 M to approximately 0.55 M), this change leads to a significant reduction in inference time (up to 16.2%) and improved classification performance. This demonstrates a favorable tradeoff between computational cost and accuracy, further validating the effectiveness of the proposed method in real-world scenarios.

Moreover, HyperEAST exhibits strong generalization capability across both small- and large-scale datasets and demonstrates resilience to severe class imbalance. In particular, on WHU-Hi-HC and WHU-Hi-LK, it achieves high accuracy with reduced inference latency. These findings highlight HyperEAST as an effective and efficient solution for real-world hyperspectral classification tasks.

F. Summary and Analysis of Experimental Results

We conducted extensive experiments on four widely used benchmark datasets—IP, PaviaU, WHU-Hi-HC, and WHU-Hi-LK—to evaluate the performance of our proposed HyperEAST framework. The results demonstrate the effectiveness of each proposed component, including LFAM, modality-decoupled MIM, and the HLoss strategy. Key findings are summarized as follows:

In addition to quantitative metrics, we provide qualitative visualizations of the classification maps produced by HyperEAST and baseline methods. As shown in Figs. 7–10, the predicted results on all four datasets are displayed alongside ground-truth and baseline transformer models. To enhance interpretability, we include zoomed-in views of key regions that highlight improvements in edge sharpness, interclass separation, and noise suppression. These visual results further validate that

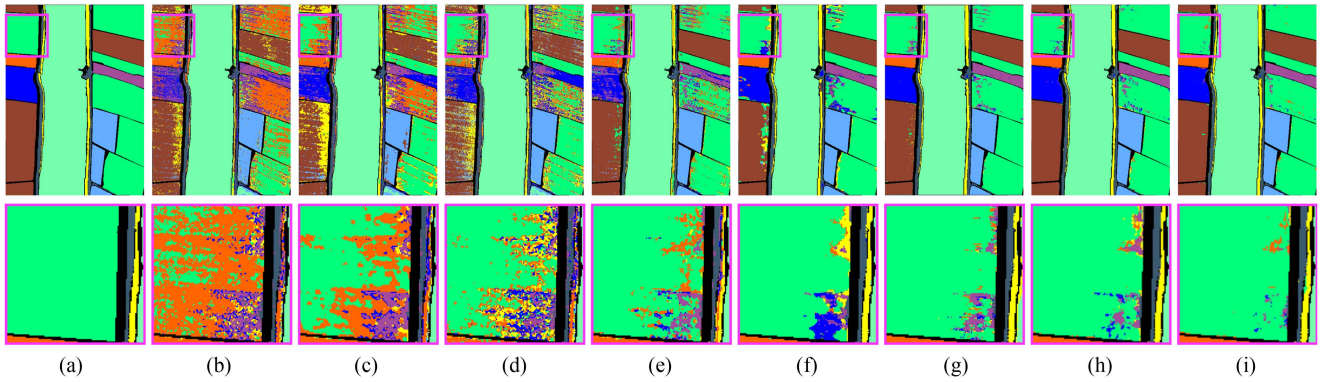

Fig. 8. Visualization of ground-truth and classification maps obtained by different models on the WHU-Hi-LK dataset (from left to right: (a) ground truth, (b) SF (pixel), (c) SF (patch), (d) MAEST (pixel), (e) MAEST (patch), (f) GAHT, (g) Swin-MSP, (h) FactoFormer, and (i) HyperEAST).

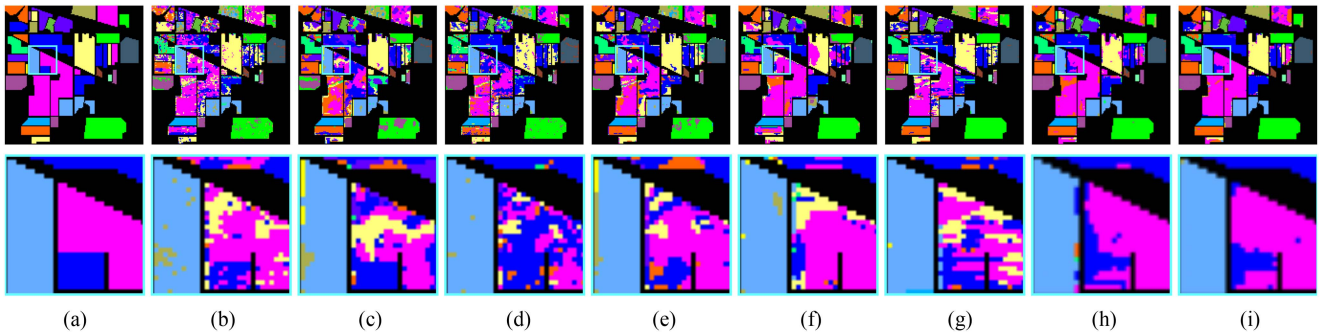

Fig. 9. Visualization of ground-truth and classification maps obtained by different models on the IP dataset (from left to read: (a) ground truth, (b) SF (pixel), (c) SF (patch), (d) MAEST (pixel), (e) MAEST (patch), (f) GAHT, (g) Swin-MSP, (h) FactoFormer, and (i) HyperEAST).

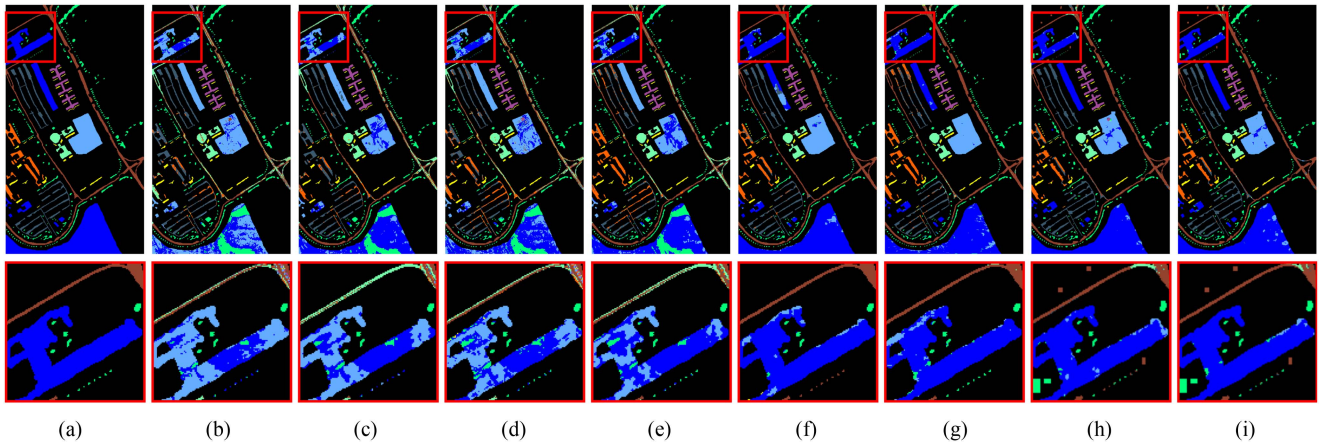

Fig. 10. Visualization of ground-truth and classification maps obtained by different models on the PaviaU dataset (from left to read: (a) ground truth, (b) SF (pixel), (c) SF (patch), (d) MAEST (pixel), (e) MAEST (patch), (f) GAHT, (g) Swin-MSP, (h) FactoFormer, and (i) HyperEAST).

HyperEAST achieves better spatial consistency and fine-grained classification performance, especially in complex or spectrally similar regions.

1) *Superior Classification Performance:* HyperEAST achieves the highest OA across all datasets compared to strong baselines, including SpectralFormer, MAEST, Swin-MSP, and Factoformer. Notably, on IP and WHU-Hi-LK, it attains 93.81 and 98.87 OA, respectively, indicating its strong generalization capability in complex scenes.
2) *Effectiveness of Self-Supervised Pretraining:* The proposed modality-decoupled MIM consistently improves model robustness under limited supervision. Pretrained models outperform randomly initialized counterparts,

particularly on IP and Houston2013, where labeled data are scarce.

3) *Efficient Attention With LFAM:* Ablation studies reveal that LFAM notably reduces computational overhead while maintaining or improving classification performance. Its grouped convolution design achieves better efficiency–accuracy tradeoff compared to standard MHSA.
4) *HLoss Improves Minority Class Learning:* Incorporating FL during fine-tuning enhances discrimination for hard or underrepresented classes. The proposed dataset-aware HLoss consistently yields higher AA and Kappa, especially on class-imbalanced datasets, such as IP.

V. Discussion

In this work, we explore the tradeoff between model complexity and performance by integrating a lightweight attention alternative—LFAM—into a dual-branch transformer architecture for HSI classification. Notably, although the use of LFAM leads to a slight increase in both parameter count and FLOPs, we observe a consistent reduction in overall computation time, especially in the spectral branch. This demonstrates that FLOPs are not always a reliable proxy for practical efficiency; LFAM replaces high-cost matrix multiplications with convolution-based additive operations that are more compatible with modern hardware, thereby accelerating inference.

The increase in parameter count and FLOPs primarily arises from replacing the MHSA with multiple additive convolutional modules. While MHSA typically uses shared linear projections for different heads, LFAM utilizes independent convolutional filters in both spatial and channel branches to capture diverse local and global patterns. This structural shift introduces more learnable parameters; however, these operations are highly parallelizable and optimized for modern GPU architectures. Crucially, despite this increase, the overall parameter count of our model remains extremely lightweight—approximately 0.5M—which is significantly lower than most transformer-based HSI models. This makes our method highly suitable for real-time or embedded applications, such as UAVs and portable HSI sensors, where fast inference and low memory footprint are essential.

Interestingly, when LFAM is applied to the spectral branch (denoted as LFAM-SPE), it not only reduces computation time but also improves classification accuracy. We hypothesize that this improvement stems from a combination of enhanced feature extraction and reduced overfitting. On one hand, the spectral domain exhibits high interband redundancy, and traditional MHSA may overfit to irrelevant fluctuations due to its dense pairwise modeling. On the other hand, LFAM's use of convolutional operations—particularly the parallel spatial and channel branches—introduces useful inductive biases that help regularize the model while also enhancing its ability to extract informative local spectral patterns. In addition, the sequential nature of spectral bands is preserved without the need for positional encoding, further contributing to modeling stability.

However, when LFAM is deployed in the spatial branch (LFAM-SPA), the benefits diminish or even reverse—computation time increases and classification accuracy drops. This can be attributed to the spatial domain's demand for long-range context modeling and fine-grained structural representation, such as edges, textures, and object boundaries. The localized nature of LFAM's convolutional paths lacks the dynamic, global receptive field that MHSA naturally offers. Furthermore, in high-resolution spatial representations, convolution operations may become computational bottlenecks due to increased spatial dimensions, offsetting LFAM's theoretical runtime gains.

These findings highlight a critical insight: lightweight attention mechanisms, such as LFAM, should be deployed selectively, with attention to the modality-specific characteristics of the data. While LFAM proves to be highly effective in the spectral branch, its blanket application to all branches may lead to suboptimal outcomes. Future work may explore hybrid strategies that combine the efficiency of LFAM with the expressiveness of MHSA in a complementary manner, potentially through adaptive attention routing based on modality or input characteristics. In addition to efficiency and accuracy, robustness is an essential aspect of real-world HSI classification. HSIs are often affected by noise, sensor malfunctions, or missing bands caused by atmospheric absorption or calibration errors. Although our current evaluation is conducted on clean benchmark datasets, HyperEAST demonstrates intrinsic robustness to such corruptions due to the use of modality-aware MIM during pretraining. Specifically, we employ a 70% masking ratio on both spectral and spatial tokens, encouraging the model to reconstruct missing or degraded information using surrounding context. This pretraining strategy enhances the model's capability for spectral inpainting and denoising, leading to improved resilience against corrupted inputs during inference. In future work, we aim to further evaluate this robustness by explicitly introducing band-level noise or simulating real-world sensor degradations to quantify the gains achieved through MIM-based pretraining.

VI. Limitations and Generalization

While HyperEAST achieves competitive performance in HSI classification, several limitations remain. First, although an HLoss is employed to mitigate class imbalance, the model may still struggle to generalize on extremely underrepresented classes due to limited feature diversity. Second, the current fusion strategy—simple token concatenation—lacks explicit spectral–spatial interaction, potentially underutilizing modality complementarity. Third, the proposed masked pretraining strategy relies on access to sufficient unlabeled data; in data-scarce domains, the benefits of self-supervision may diminish.

Moreover, while HyperEAST has been validated on several benchmark datasets with diverse characteristics, its generalization to other HSI sensors—such as satellite-based platforms or data with different spectral resolutions—remains an open question. Sensor-specific factors, including noise patterns and geospatial variability, may affect performance. Future work may explore domain adaptation techniques, sensor-aware pretraining, and enhanced fusion modules to improve robustness and cross-sensor transferability.

VII. Conclusion

We present *HyperEAST*, a novel dual-branch ViT framework tailored for HSI classification. HyperEAST addresses key challenges in HSI analysis, including high spectral dimensionality, limited labeled data, and severe class imbalance. The model explicitly decouples spectral and spatial representation learning through a two-branch architecture, enabling specialized modeling of long-range spectral dependencies and localized spatial patterns.

To improve inference efficiency, we adopt the LFAM—a convolutional additive design that serves as a lightweight alternative to conventional dot-product attention. By replacing MHSA with convolutional operations in the spectral branch, LFAM facilitates more efficient and stable representation learning, achieving improved accuracy and reduced latency.

To mitigate reliance on annotated data, we adopt a self-supervised MIM strategy for spectral–spatial representation pretraining. During fine-tuning, we employ an HLoss that combines CE and FL to address class imbalance and improve robustness.

Extensive experiments on four benchmark datasets confirm that HyperEAST achieves competitive performance in terms of OA, AA, and Kappa coefficient, while also reducing inference latency. In summary, HyperEAST offers a scalable, label-efficient, and deployment-ready solution for practical HSI classification, making it well-suited for real-world applications in remote sensing and geospatial data mining.

References

[1] X. Zhang et al., "SSDANet: Spectral-spatial three-dimensional convolutional neural network for hyperspectral image classification," *IEEE Access*, vol. 8, pp. 127167–127180, 2020.

[2] P. Ranjan, A. Girdhar Ankur, and R. Kumar, "A novel spectral-spatial 3D auxiliary conditional GAN integrated convolutional LSTM for hyperspectral image classification," *Earth Sci. Informat.*, vol. 17, pp. 5251–5271, 2024.

[3] J. Wu, X. Sun, L. Qu, X. Tian, and G.-Y. Yang, "Learning spatial–spectral-dimensional-transformation-based features for hyperspectral image classification," *Appl. Sci.*, vol. 13, no. 14, 2023, Art. no. 8451.

[4] K. Mounika, K. Aravind, M. Yamini, P. Navyasri, S. Dash, and V. Suryanarayana, "Hyperspectral image classification using SVM with PCA," in *Proc. 6th Int. Conf. Signal Process., Comput. Control*, Solan, India, 2021, pp. 470–475.

[5] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, 2017.

[6] B. Liu, X. Yu, P. Zhang, X. Tan, A. Yu, and Z. Xue, "A semi-supervised convolutional neural network for hyperspectral image classification," *Remote Sens. Lett.*, vol. 8, no. 9, pp. 839–848, 2017.

[7] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3D–2D 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.

[8] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 5998–6008.

[9] D. Ibanez, R. Fernandez-Beltran, F. Pla, and N. Yokoya, "Masked auto encoding spectral–spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5542614.

[10] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[11] X. Fang, G. Zhang, G. Zhang, X. Zhou, J. Wu, and L. Zhao, "A hybrid self-supervised learning framework for hyperspectral image classification," in *Proc. 2023 Int. Conf. Comput., Vis. Intell. Technol.*, 2023, pp. 1–7.

[12] L. Tu, J. Li, X. Huang, J. Gong, X. Xie, and L. Wang, "S2HM2: A Spectral–Spatial Hierarchical Masked Modeling Framework for Self-Supervised Feature Learning and Classification of Large-Scale Hyperspectral Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5517019.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[14] B. Ma, H. Wang, and L. Wang, "A model based on dense connection 3D2DCNN for hyperspectral image classification," in *Proc. 5th Int. Conf. Video, Signal Image Process.*, 2023, pp. 37–42, doi: 10.1145/3638682.3638688.

[15] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[16] K. M. Lim, C. P. Lee, Z. Zahisham, J. Y. Lim, and J. N. Mogan, "PCA-ViT: Hyperspectral image classification using principal component analysis and vision transformer," in *Proc. IEEE 12th Conf. Syst., Process Control*, 2024, pp. 30–34.

[17] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539014.

[18] Y. Tan, M. Li, L. Yuan, C. Shi, Y. Luo, and G. Wen, "Hyperspectral image classification with embedded linear vision transformer," *Earth Sci. Informat.*, vol. 18, 2024, Art. no. 69.

[19] C. Tao, J. Qi, M. Guo, Q. Zhu, and H. Li, "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610426, doi: 10.1109/TGRS.2023.3276853.

[20] Y. Guo, Q. Yu, Y. Gao, X. Liu, and C. Li, "Max–min distance embedding for unsupervised hyperspectral image classification in the satellite Internet of Things," *Internet Things*, vol. 22, 2023, Art. no. 100775.

[21] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.

[22] J. Lin, F. Gao, X. Shi, J. Dong, and Q. Du, "SS-MAE: Spatial–spectral masked autoencoder for multisource remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531614.

[23] L. Zhu, J. Wu, B. Wang, Y. Liao, and D. Gu, "SpectralMAE: Spectral Masked Autoencoder for Hyperspectral Remote Sensing Image Reconstruction," *Sensors*, vol. 23, no. 7, Apr. 2023, doi: 10.3390/s23073728, Art. no. 3728.

[24] M. Ahmad, M. Mazzara, S. Distefano, A. M. Khan, and X. Wu, "Self-supervised spatial-spectral transformer with Extreme Learning Machine for hyperspectral image classification," *Proc. Int. J. Remote Sens.*, vol. 46, no. 14, Jul. 2025, pp. 1–24, doi: 10.1080/01431161.2025.2520049.

[25] B. Zhao, Y. Qin, and M. Yang, "Convolution-enhanced vision transformer for hyperspectral image classification," in *Proc. Int. Geosci. Remote Sens. Symp. 2024*, Athens, Greece, 2024, pp. 8820–8824.

[26] S. Arya, S. R. Dubey, S. M. Moorthi, D. Dhar, and S. K. Singh, "3D-CmT: 3D-CNN meets transformer for hyperspectral image classification," in *∗Proc. Asian Conf. Comput. Vis. (ACCV) Workshops∗*, Dec. 2024, pp. 679–695.

[27] M. Ahmad, M. H. F. Butt, M. Mazzara, S. Distefano, A. M. Khan, and H. A. Altuwaijri, "Pyramid hierarchical spatial-spectral transformer for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 17681–17689, 2024, doi: 10.1109/JSTARS.2024.3461851.

[28] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[29] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: arXiv:2110.02178

[30] R. Tian, D. Liu, Y. Bai, Y. Jin, G. Wan, and Y. Guo, "Swin-MSP: A shifted windows masked spectral pretraining model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4509114.

[31] Y. Li, Y. Luo, L. Zhang, Z. Wang, and B. Du, "MambaHSI: Spatial–spectral mamba for hyperspectral image classification," in *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5524216, doi: 10.1109/TGRS.2024.3430985.

[32] S. Mohamed, M. Haghighat, T. Fernando, S. Sridharan, C. Fookes, and P. Moghadam, "FactoFormer: Factorized hyperspectral transformers with self-supervised pretraining," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5501614.

[33] X. Yang, W. Cao, D. Tang, Y. Zhou, and Y. Lu, "ACTN: Adaptive coupling transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5503115.

[34] S. Varahagiri, A. Sinha, S. R. Dubey, and S. Kuamr Singh, "3D-Convolution guided spectral-spatial transformer for hyperspectral image classification," in *Proc. IEEE Conf. Artif. Intell.*, Singapore, 2024, pp. 8–14, doi: 10.1109/CAI59869.2024.00011.
[35] T. Zhang, L. Li, Y. Zhou, W. Liu, C. Qian, J.-N. Hwang, and X. Ji, "CAS-ViT: Convolutional additive self-attention vision transformers for efficient mobile applications," Aug. 2024, *arXiv:2408.03703*.
[36] L. Scheibenreif, M. Mommert, and D. Borth, "Masked Vision Transformers for Hyperspectral Image Classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Vancouver, BC, Canada, 2023, pp. 2166–2176, doi: 10.1109/CVPRW59228.2023.00210.
[37] K. Li, Y. Chen, and L. Huang, "Dual-branch masked transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5510805, doi: 10.1109/LGRS.2024.3490534.
[38] J. Jiao, C. Yin, and F. Teng, "W-net: Deep convolutional network with gray-level co-occurrence matrix and hybrid loss function for hyperspectral image classification," ∗*Adv. Intell. Comput. Technol. Appl. (ICIC 2023)*∗, Springer, 2023, pp. 112–124.
[39] W. Zhang, M. Hu, S. Hou, R. Shang, J. Feng, and S. Xu, "Group-spectral superposition and position self-attention transformer for hyperspectral image classification," ∗*Expert Syst. Appl.*∗, 2024, Art. no. 125846, doi: 10.1016/j.eswa.2024.125846.
[40] Y. Fang, L. Sun, Y. Zheng, and Z. Wu, "Deformable convolution-enhanced hierarchical transformer with spectral-spatial cluster attention for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 34, pp. 701–716, 2025, doi: 10.1109/TIP.2024.3522809.
[41] X. Su and J. Shao, "3DVT: Hyperspectral image classification using 3D dilated convolution and mean transformer," *Photon.*, vol. 12, no. 2, Feb. 2025, doi: 10.3390/photonics12020146.
[42] B. Vaishnavi, A. Pamidighantam, A. Hema, and V. R. Syam, "Hyperspectral image classification for agricultural applications," in *Proc. 2022 Int. Conf. Electron. Renewable Syst.*, 2022, pp. 1–7.
[43] H. Chen, F. Miao, Y. Chen, Y. Xiong, and T. Chen, "A hyperspectral image classification method using multifeature vectors and optimized KELM," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 2781–2795, 2021.
[44] C. Wang, J. Huang, M. Lv, H. Du, Y. Wu, and R. Qin, "A local enhanced mamba network for hyperspectral image classification," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 133, 2024, Art. no. 104092.
[45] H. Huang, C. Pu, Y. Li, and Y. Duan, "Adaptive residual convolutional neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2520–2531, 2020.
[46] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539014.

Jialin Tang received the B.S. degree in information management and information systems from the Shandong University of Finance and Economics, Jinan, China, in 2022. He is currently working toward the M.S. degree in computer science with the College of Engineering and Computer Science, California State University Fullerton, Fullerton, CA, USA.

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

Nan Ma received the D.Eng. degree in surveying and mapping engineering from the China University of Petroleum, Qingdao, China, in 2023.

She is currently an Associate Professor with the School of Artificial Intelligence, Shandong Women's University, Jinan, China. Her research interests include remote sensing and computer vision.

Chen Jia received the B.S. degree in mathematics and the M.S. and Ph.D. degrees in remote sensing from the Shandong University of Science and Technology, Qingdao, China, in 2014, 2017, and 2024, respectively.

He is currently a Professional Teacher with the School of Artificial Intelligence, Shandong Women's University, Jinan, China. His research interests include the retrieval of aerosol optical depth from satellite data and the application of deep learning in remote sensing.

Rui Tian received the B.S. degree in computer science and technology from the School of Computer Science and Technology, Shandong Jianzhu University, Jinan, China, in 2019. He is currently working toward the M.Sc. degree in electronic information with the College of Computer, Qinghai Normal University, Qinghai, China.

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

Yanhui Guo (Member, IEEE) received the B.S. degree in information management and information system from the Xi'an University of Finance and Economics, Xi'an, China, in 2006, and the M.S. and Ph.D. degrees in computer software and theory from Shaanxi Normal University, Xi'an China, in 2009 and 2020, respectively.

Since 2009, he has been with the the School of Artificial Intelligence, Shandong Women's University, Ji'nan, China. He is currently a Professor. His research interests include machine learning and remote sensing image processing.