


DiffFormer: A Differential Spatial-Spectral Transformer for Hyperspectral Image Classification

Muhammad Ahmad , Manuel Mazzara , Salvatore Distefano, Adil Mehmood Khan ,
and Silvia Liberata Ullo , *Senior Member, IEEE*

Abstract—Hyperspectral image classification (HSIC) presents significant challenges due to spectral redundancy and spatial discontinuity, both of which can negatively impact classification performance. To mitigate these issues, this work proposes the differential spatial-spectral transformer (*DiffFormer*), a novel framework designed to enhance feature discrimination and improve classification accuracy. At its core, *DiffFormer* incorporates a differential multihead self-attention mechanism, which accentuates subtle spectral-spatial variations by applying differential attention across neighboring patches. The architecture integrates spectral-spatial tokenization, utilizing 3-D convolution-based patch embeddings, positional encoding, and a stack of transformer layers augmented with the SwiGLU activation function—a variant of the gated linear unit—to enable efficient and expressive feature extraction. In addition, a token-based classification head ensures robust representation learning, facilitating precise pixelwise labeling. Extensive experiments on benchmark hyperspectral datasets demonstrate that *DiffFormer* consistently outperforms state-of-the-art methods in classification accuracy, computational efficiency, and generalizability.

Index Terms—Differential attention, hyperspectral image classification (HSIC), spatial-spectral transformer (SST).

I. INTRODUCTION

HYPERSPECTRAL imaging has emerged as a transformative technology with diverse applications, including precision agriculture [1], object classification [2], environmental monitoring [3], urban mapping for mineral exploration [4], [5],

food processing [6], [7], bakery products [8], bloodstain identification [9], [10], and meat processing [11], [12]. By capturing detailed spectral information at the pixel level, hyperspectral image (HSI) enables fine-grained material classification [13]. However, its practical deployment is constrained by challenges such as high dimensionality, spectral variability, complex spatial structures, and the Hughes phenomenon [14], [15]. Addressing these challenges requires classification frameworks that are not only highly accurate but also computationally efficient and capable of effectively integrating spatial-spectral information [16], [17], [18], [19].

Recent advancements in Transformer-based models have demonstrated significant success in hyperspectral image classification (HSIC) by leveraging self-attention mechanisms to process image patches effectively [20], [21], [22], [23], [24], [25]. For example, Huang et al. [26] introduced the spectral-spatial vision foundation model-based transformer (SS-VFMT), which enhances pretrained vision foundation models (VFMs) with specialized spectral and spatial enhancement modules. They further propose a patch relationship distillation strategy (SS-VFMT-D) to optimize the utilization of pretrained knowledge and introduce the spectral-spatial vision-language transformer (SS-VLFMT) for generalized zero-shot classification, enabling the recognition of previously unseen classes. Although these methods achieve impressive performance, they are constrained by high-computational complexity and a strong dependence on pretrained models, which may limit their adaptability to domain-specific applications.

Shu et al. [27] proposed a dual attention transformer network (DATN) for HSIC that integrates a spatial-spectral hybrid transformer module to capture global spatial-spectral dependencies and a spectral local-conv block (SLCB) module to extract local spectral features effectively. While DATN enhances feature representation by combining global and local information, its reliance on multihead self-attention (MHSA) mechanisms may still lead to computational overhead. Zhong et al. [28] proposed a spectral-spatial transformer network that integrates spatial attention and spectral association modules to address convolutional limitations while a factorized architecture search framework enables efficient architecture optimization. Despite achieving competitive accuracy with reduced computational costs, relying on a specialized architecture search framework may limit flexibility for broader applications.

Yang et al. [29] proposed the quaternion transformer network, which addresses the limitations of traditional

Received 20 February 2025; revised 24 March 2025; accepted 3 April 2025. Date of publication 8 April 2025; date of current version 25 April 2025. This work was supported in part by the European Union - Next generation EU - PNRR - Missione 4, Componente 2, Investimento 1.1 - Bando PRIN 2022 PNRR - Decreto Direttoriale n. 1409 del 14-09-2022 - Progetto RESILIENT, under Grant CUP J53D23015040001, project id. P2022S4TTP, and in part by the Next generation EU - PNRR, SERICS – “SECURITY AND RIGHTS IN THE CYBERSPACE” project 3D-SEECSD, under Grant CUP J33C22002810001, project id. PE00000014. (Corresponding author: Muhammad Ahmad.)

Muhammad Ahmad is with the SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRCAI), King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia (e-mail: mahmad00@gmail.com).

Manuel Mazzara is with the Institute of Software Development and Engineering, Innopolis University, 420500 Innopolis, Russia (e-mail: m.mazzara@innopolis.ru).

Salvatore Distefano is with Dipartimento di Matematica e Informatica-MIFT, University of Messina, 98121 Messina, Italy (e-mail: sdistefano@unime.it).

Adil Mehmood Khan is with the School of Computer Science, University of Hull, HU6 7RX Hull, U.K. (e-mail: a.m.khan@hull.ac.uk).

Silvia Liberata Ullo is with the Engineering Department, University of Sannio, 82100 Benevento, Italy (e-mail: mahmad00@gmail.com).

<https://github.com/mahmad000/DiffFormer>.

Digital Object Identifier 10.1109/JSTARS.2025.3558889

transformers in HSIC by leveraging quaternion algebra for efficient 3-D structure processing and spectral-spatial representation. However, the reliance on hypercomplex computations may increase implementation complexity and resource requirements. Zhang et al. [30] proposed a LiT network that integrates lightweight self-attention modules and convolutional tokenization to balance local feature extraction and global dependency capture for HSIC. Despite improved efficiency and reduced overfitting, its reliance on controlled sampling strategies may limit adaptability to diverse datasets. Yang et al. [31] proposed a HiT classification network that combines spectral-adaptive 3-D convolution and Conv-Permutator modules to enhance spatial-spectral representation in HSIs, addressing convolutional neural networks (CNNs)' limitations in mining spectral sequences. However, the added complexity of these modules may increase computational overhead. Yu et al. [32] proposed the MST-Net combining a self-attentive encoder and multilevel features decoding within an efficient transformer-based framework for HSIC. However, reliance on sequence-based processing and positional embeddings may underexploit the inherent 3-D structure of HSIs. Zhang et al. [33] proposed the MATNet which integrates multiattention mechanisms and transformers for HSIC, improving boundary pixel classification with spatial-channel attention and a novel Lpoly loss function. However, reliance on semantic-level tokenization may limit fine-grained feature preservation.

Ye et al. [34] introduced a novel differential transformer architecture designed for noise cancellation. However, its applicability to HSIC is constrained by several limitations. First, the reliance on subtractive attention for sparse patterns may fail to effectively address the spectral redundancy and spatial discontinuity inherent in HSIC, resulting in suboptimal feature discrimination. Furthermore, the model is primarily designed for text-based tasks and lacks essential components such as spectral-spatial tokenization and domain-specific architectural adaptations required for high-dimensional remote sensing data, which are fundamental to spatial-spectral transformers (SSTs).

Moreover, SSTs themselves present several challenges for instance, training large-scale SST models is computationally expensive, primarily due to the quadratic complexity of the self-attention mechanism with respect to sequence length, which limits scalability [35], [36], [37]. Unlike CNNs, which inherently encode translation invariance through shared-weight convolutional filters, SSTs often struggle to maintain robust spatial representations under minor input translations [38], [39], [40]. In addition, the fixed-size patch tokenization employed in SSTs may hinder the model's ability to capture fine-grained spectral-spatial details, thereby affecting classification performance [41], [42], [43]. Another critical limitation is the requirement for large labeled datasets to achieve optimal performance. In data-limited scenarios, SSTs are prone to overfitting, restricting their effectiveness in real-world applications with limited annotated hyperspectral samples [44]. Therefore, this study proposes a novel framework designed to overcome the aforementioned limitations through the following key contributions.

- 1) *Differential Attention Mechanism for Localized Feature Discrimination*: The proposed differential MHSA

(DMHSA) mechanism enhances hyperspectral feature representation by computing differential attention scores. These scores highlight subtle spectral-spatial variations while mitigating spectral redundancy and spatial noise, allowing for more discriminative interpixel modeling. This approach significantly improves classification accuracy by refining spectral-spatial feature dependencies.

- 2) *Integration of SwiGLU Activation in Spectral-Spatial Transformers*: The differential spatial-spectral transformer (*DiffFormer*) integrates SwiGLU, a variant of the gated linear unit (GLU) activation function, into the feed-forward layers of the transformer blocks. SwiGLU's adaptive gating mechanism enables selective feature enhancement, improving the model's capacity to capture intricate spectral-spatial dependencies while maintaining computational efficiency.
- 3) *Class Token for Unified Spectral-Spatial Representation*: To achieve global spectral-spatial feature aggregation, a learnable class token is introduced, summarizing information across hyperspectral bands. In addition, sinusoidal positional encoding is applied to hyperspectral data, ensuring precise spatial and spectral continuity modeling. This tailored encoding enhances the alignment between spectral-spatial tokenization and feature representation, a critical factor for improving HSIC performance.
- 4) *Efficient Patch-Based Spectral-Spatial Tokenization*: The *DiffFormer* employs a 3-D convolutional patch embedding strategy that simultaneously extracts spectral and spatial features while reducing input dimensionality. This method ensures that essential spectral-spatial relationships are preserved, allowing for computationally efficient processing without sacrificing feature richness.

II. PROPOSED METHODOLOGY

This section introduces the proposed differential SST (*DiffFormer*) model for HSIC. *DiffFormer* integrates differential attention mechanisms and spatial-spectral tokenization to effectively encode both spatial and spectral dependencies. The architectural components and key innovations are illustrated in Fig. 1. Designed to capture complex spatial and spectral relationships in HSIs, *DiffFormer* combines convolutional and transformer-based methodologies. Its core components include spatial-spectral patch embedding, positional encoding, the DMHSA module, and a classification head.

A. Spatial-Spectral Patch Embedding

An HSI cube $\mathcal{X} = \{x_i, y_i\} \in \mathbb{R}^{H \times W \times K}$ consists of spectral vectors $x_i = \{x_{i,1}, x_{i,2}, x_{i,3}, \dots, x_{i,K}\}$, where x_i represents the spectral information of pixel i , and y_i denotes its corresponding class label. Here, H and W define the spatial dimensions (height and width) of the HSI, while K represents the number of spectral channels. Each pixel i is described by a spectral vector $x_i \in \mathbb{R}^k$, with $\{x_{i,1}, x_{i,2}, \dots, x_{i,K}\}$ representing its spectral values, and y_i being an integer indicating the class assignment.

The HSI cube is partitioned into overlapping 3-D patches, where each patch is centered at spatial coordinates (α, β) and

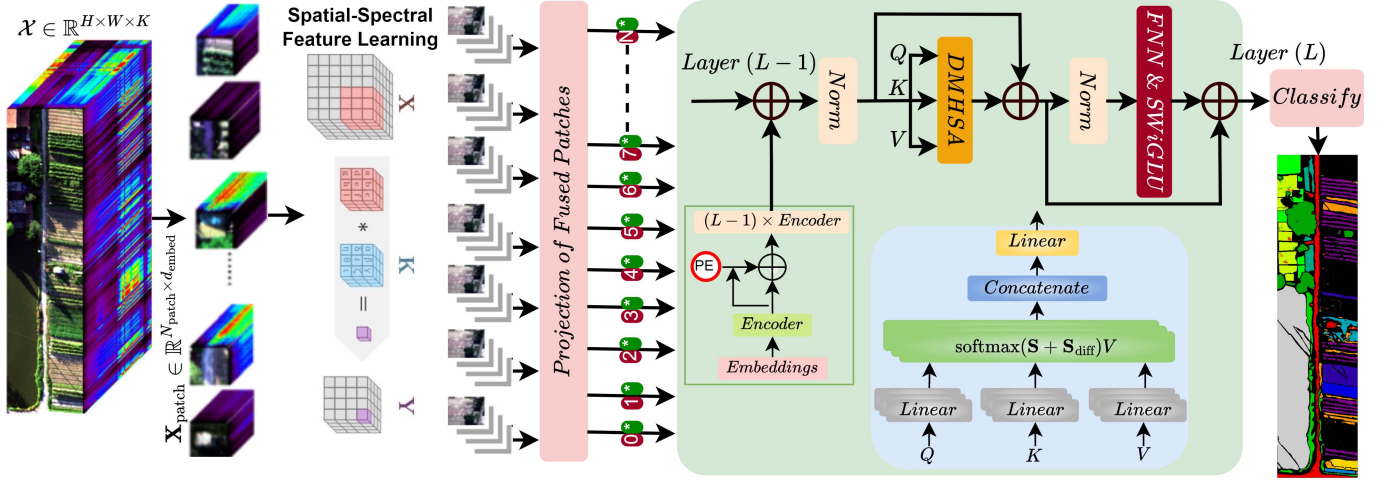


Fig. 1. Schematic representation of the *DiffFormer* pipeline for HSIC. The pipeline starts with hyperspectral data preprocessing, where fused patches are generated and spatial-spectral features are extracted. DMHSA is employed within the encoder to refine the attention mechanism by integrating positional embeddings for enhanced spectral-spatial relationships. The hierarchical encoding layers aggregate the learned features across $L - 1$ layers, enabling multiscale representation learning.

spans $P \times P$ pixels across K spectral bands. This results in a 3-D subregion that encapsulates both spatial and spectral information. The total number of extracted patches, m , is computed as follows:

$$m = (H - P + 1) \times (W - P + 1) \quad (1)$$

assuming a stride of $s = 1$. If the stride s is smaller than the patch size P , the patches overlap. The overlap ratio, r , is given as follows:

$$r = 1 - \frac{s}{P} \quad (2)$$

where r quantifies the extent of overlap between adjacent patches. When $s = P$, no overlap occurs ($r = 0$), whereas, for $s < P$, the patches overlap, with the overlap ratio increasing as the stride decreases. Each patch undergoes spatial-spectral patch embedding. A 3-D convolutional layer extracts feature representations using patches of size $P \times P \times K$, ensuring efficient spectral-spatial feature learning as follows:

$$\mathbf{X}_{\text{patch}} \in \mathbb{R}^{N_{\text{patch}} \times d_{\text{embed}}} \quad (3)$$

where N_{patch} is the number of patches and d_{embed} is the embedding dimension. In addition, a trainable class token is appended to this sequence to facilitate global feature aggregation.

B. Positional Encoding

To preserve spatial context, a sinusoidal positional encoding is incorporated into the patch embeddings as follows:

$$\mathbf{X}_{\text{pos}} = \mathbf{X}_{\text{patch}} + \mathbf{PE} \quad (4)$$

where $\mathbf{PE} \in \mathbb{R}^{(N_{\text{patch}}+1) \times d_{\text{embed}}}$ encodes spatial relationships. The positional encoding matrix is precomputed using trigonometric functions as follows:

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{embed}}}}\right) \quad (5)$$

$$\text{PE}(\text{pos}, 2i + 1) = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{embed}}}}\right). \quad (6)$$

In contrast to learnable positional embeddings, sinusoidal encoding facilitates smooth interpolation for unseen input lengths, thereby preserving spatial coherence in hyperspectral data.

C. Differential Multihead Self-Attention (DMHSA)

The core innovation of *DiffFormer* lies in the DMHSA module. Unlike conventional MHSA, which primarily captures global dependencies, DMHSA introduces differential operations on attention scores to model fine-grained variations—crucial for HSIC. The standard attention scores are computed as follows:

$$\mathbf{S} = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_{\text{head}}}} \quad (7)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ represent the query, key, and value matrices, respectively, and d_{head} is the head dimension. To introduce differential attention, the DMHSA module computes as follows:

$$\mathbf{S}_{\text{diff}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T - \Delta(\mathbf{Q}\mathbf{K}^T)}{\sqrt{d_{\text{head}}}}\right) \quad (8)$$

where $\Delta(\mathbf{Q}\mathbf{K}^T)$ captures the difference in attention scores across neighboring spectral-spatial tokens. This formulation explicitly encodes local contrastive dependencies, distinguishing spectral features that are critical for fine-grained classification.

The differential operation \mathbf{S}_{diff} introduces several key benefits: 1) Relative attention dynamics: Unlike absolute attention scores, \mathbf{S}_{diff} emphasizes changes in attention values, making it effective in detecting transitions between consecutive tokens. 2) Enhanced sensitivity to local variations: By highlighting significant attention shifts, it improves the model's ability to recognize fine-grained contextual differences—essential for HSIC tasks. 3) Sparsity and interpretability: By suppressing uniform or redundant attention patterns, the differential operation yields sparse and interpretable attention maps, reducing computational

overhead. 4) Robustness to noise: The differential formulation mitigates sensitivity to small perturbations in query-key interactions, making the model resilient in noisy environments. 5) Sequential dependency encoding: By focusing on attention transitions, DMHSA naturally captures sequential dependencies, beneficial for tasks with strong spatial or temporal correlations. 6) Multiscale representation: When combined with hierarchical or multihead attention, the mechanism effectively captures both absolute and relative importance across different granularities. 7) Regularization effect: The differential operation helps smooth attention transitions, preventing abrupt changes that could arise from overfitting, thereby enhancing model generalization. Finally, the resulting attention weights are used to compute the weighted sum of values as follows:

$$\mathbf{Z} = \mathbf{S}_{\text{diff}} \mathbf{V}. \quad (9)$$

D. Transformer Encoder Block

Each encoder block integrates DMHSA with the SwiGLU activation function to enhance feature expressiveness. The SwiGLU activation is defined as follows:

$$\text{SwiGLU}(x, g) = x \odot \sigma(g) + x \quad (10)$$

where $\sigma(\cdot)$ denotes the sigmoid function. Layer normalization and residual connections ensure stable training and effective gradient flow. The output from the final SST layer is projected through a dense layer to extract the class token representation. This class token is then processed by fully connected layers with L2 regularization to generate the classification logits as follows:

$$\mathbf{y} = \text{softmax}(\mathbf{W}\mathbf{z}_{\text{cls}} + \mathbf{b}) \quad (11)$$

where \mathbf{z}_{cls} represents the class token embedding and \mathbf{W}, \mathbf{b} are trainable parameters.

E. Computational Complexity and Implementation

To ensure a fair computational cost comparison, we analyze the complexity of *DiffFormer* relative to both standard transformer-based and CNN-based architectures. The core computational component, the DMHSA mechanism, retains the $\mathcal{O}(N_{\text{patch}}^2 \cdot d_{\text{head}})$ complexity of conventional self-attention. However, it introduces computational sparsity by leveraging spectral-spatial differential scoring, which refines attention maps and reduces redundant operations.

a) *Comparison with CNN-based architectures*: Unlike CNNs, which exhibit $\mathcal{O}(N_{\text{patch}} \cdot k^2 \cdot C^2)$ complexity (where k is the kernel size and C is the number of channels), *DiffFormer* effectively balances global spectral dependence modeling and localized feature extraction with a comparable parameter footprint. This enables better long-range spatial-spectral interactions while maintaining computational efficiency.

b) *Computational cost of DMHSA versus standard MHSA*: The standard MHSA mechanism performs full pairwise attention computation with complexity as follows:

$$\mathcal{O}_{\text{MHSA}} = \mathcal{O}(N_{\text{patch}}^2 \cdot d_{\text{head}}). \quad (12)$$

TABLE I
SUMMARY OF THE HSI DATASETS USED FOR EXPERIMENTAL EVALUATION

	SA	UH	PU	HC
Sensor	AVIRIS	CASI	ROSIS-03	Headwall Nano
Wavelength	350 – 1050	350 – 1050	430 – 860	400 – 1000
Resolution	3.7m	2.5m	1.3m	0.109m
Spatial	512 × 217	340 × 1905	610 × 610	1217 × 303
Spectral	224	144	103	274
Classes	16	15	9	16
Source	Aerial	Aerial	Aerial	Aerial

In contrast, DMHSA introduces differential scoring, modifying attention logits as follows:

$$\mathbf{S}' = \mathbf{S} + \lambda(\mathbf{S}[:, 1:] - \mathbf{S}[:, -1]) \quad (13)$$

where \mathbf{S} is the self-attention score matrix. This results in an additional term of $\mathcal{O}(N_{\text{patch}} \cdot d_{\text{head}})$, yielding an overall complexity of

$$\mathcal{O}_{\text{DMHSA}} = \mathcal{O}(N_{\text{patch}}^2 \cdot d_{\text{head}}) + \mathcal{O}(N_{\text{patch}} \cdot d_{\text{head}}). \quad (14)$$

Since $N_{\text{patch}} \gg 1$ in HSI, the additional term remains asymptotically insignificant, leading to a negligible increase in computational overhead (4.6%), as verified in our empirical runtime analysis.

c) *Implementation and Experimental Setup*: All experiments are conducted using TensorFlow with mixed-precision training and GPU acceleration for computational efficiency. We utilize NVIDIA RTX 3090 GPUs for training, employing the Adam optimizer with a learning rate of 10^{-4} . Performance is evaluated based on the following metrics overall accuracy (OA), average accuracy (AA), and Kappa coefficient (κ). To ensure reproducibility, we use identical random seeds, dataset partitions, and training hyperparameters across all comparative models.

III. EXPERIMENTAL DATASETS AND SETTINGS

To validate the efficacy of the *DiffFormer*, we utilize four HSI datasets characterized by diverse spatial and spectral features. This section details the datasets, experimental setup, evaluation metrics, and comparative results to demonstrate the robustness and adaptability of the *DiffFormer*. The experimental datasets include WHU-Hi-HanChuan (HC) [45], Salinas (SA), Pavia University (PU), and University of Houston (UH). Table I provides the experimental datasets' key details and characteristics.

The experimental settings for evaluating the *DiffFormer* were designed to ensure a robust assessment of its performance across the different datasets. During the training phase, the Adam optimizer from the TensorFlow legacy module is used with a learning rate of 0.001 and a decay rate of 1×10^{-6} , helping the model converge efficiently while minimizing the risk of overfitting. The training process spans 50 epochs, with a batch size of 56, which is large enough to allow for efficient gradient updates without overwhelming the system's memory capacity. The feed-forward network within each transformer layer consists of 4×64 units, enabling the model to extract rich and complex features from the input data. To mitigate overfitting, dropout is applied with a rate of 0.1 and layer normalization is incorporated with an epsilon value of 1×10^{-3} to stabilize the learning process. In addition,

TABLE II
CLASSIFICATION PERFORMANCE ACROSS DIFFERENT PATCH SIZES ON FOUR DATASETS, WITH THE HIGHEST VALUES FOR EACH METRIC HIGHLIGHTED IN BOLD

Patch	HC			UH			SA			PU		
	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA
8 × 8	95.20	95.90	91.27	97.29	97.49	96.88	98.20	98.38	99.23	96.87	97.64	96.58
10 × 10	96.56	97.06	93.65	96.81	97.05	96.03	98.76	98.89	99.41	97.54	98.15	97.35
12 × 12	96.34	96.87	93.55	97.38	97.57	97.08	99.13	99.22	99.55	96.76	97.56	96.90
14 × 14	97.09	97.52	95.29	97.39	97.59	96.76	99.06	99.15	99.48	98.00	98.49	97.41
16 × 16	96.50	97.01	92.48	96.61	96.87	96.29	99.53	99.58	99.76	97.37	98.01	97.54
18 × 18	97.26	97.66	95.65	80.85	82.31	77.89	99.48	99.54	99.77	97.98	98.48	97.65
20 × 20	96.56	97.06	93.65	97.91	98.07	97.90	99.06	99.16	99.16	97.77	98.32	97.25

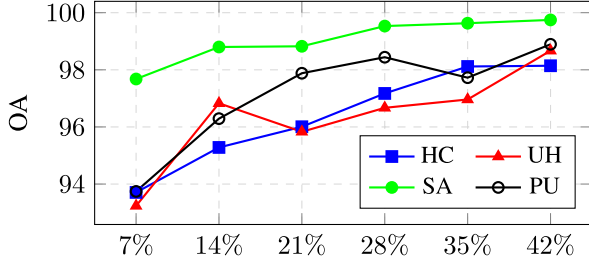


Fig. 2. Classification performance of different percentages of training samples.

a kernel regularizer with an L2 penalty of 0.01 is applied to the model's weights to further promote generalization and avoid overfitting. To reduce the dimensionality of the spectral data, PCA is employed, selecting the top 15 spectral bands that contribute most significantly to the variance, thereby retaining the most relevant information for classification.

After training and validating the model, its performance is evaluated on the test set using several key metrics: OA, which reflects the proportion of correctly classified samples across all classes; AA, which provides the mean classification accuracy per class; per-class accuracy, which evaluates the performance of the model for each class; and the κ , a measure that accounts for agreement between the predicted and true labels, adjusted for chance. These metrics together provide a comprehensive assessment of the model's effectiveness in various aspects.

IV. ABLATION AND PARAMETER OPTIMIZATION

The choice of patch size significantly impacts the performance of HSIC models by influencing the balance between spatial detail and contextual information. Smaller patches allow the model to capture fine-grained spatial features but may lack broader contextual awareness, while larger patches provide more spatial context but could blur subtle spectral distinctions. The results in Table II highlight this tradeoff, showing that larger patch sizes generally yield superior classification performance across most datasets. However, excessively large patches can lead to diminishing returns or even performance degradation in some cases, as seen with the HC and UH datasets. These findings underscore the importance of selecting an optimal patch size to maximize classification accuracy while maintaining computational efficiency.

The number of training samples directly affects the classification performance of HSIC models by influencing generalization ability and robustness. As shown in Fig. 2, increasing the percentage of training samples generally improves accuracy

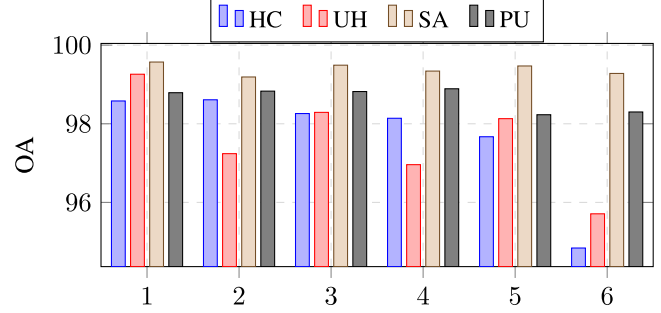


Fig. 3. Impact of transformer layer depth on HSIC.

TABLE III
IMPACT OF ATTENTION HEADS ON OA ACROSS FOUR DATASETS

Attention Heads	HC	UH	PU	SA
2	98.8079	97.6713	99.2681	99.3472
4	98.6124	97.8487	98.6597	99.4580
6	97.6805	98.3588	98.8155	99.4149
8	98.6150	99.2681	98.8311	99.5750

TABLE IV
IMPACT OF DIFFERENT ACTIVATION FUNCTIONS ON *DiffFormer*

Activation	UH			SA			PU		
	κ	OA	AA	κ	OA	AA	κ	OA	AA
ReLU	98.8487	98.9354	98.1896	99.1091	99.1489	99.2460	98.2756	98.6986	98.2626
LReLU	97.9135	98.0705	97.4061	99.6639	99.6982	99.79283	98.3689	98.7687	98.2413
PReLU	98.3046	98.3568	98.1902	98.6914	98.7228	98.8047	98.6371	98.9714	98.6163
ELU	98.3931	98.5140	97.7495	98.6091	98.6489	98.6048	98.3245	98.7376	98.3307
GELU	98.9448	99.0241	98.4808	98.6640	98.6982	98.8361	98.7078	99.0259	98.4005
SwiGLU	99.2087	99.2681	99.1277	99.5269	99.5750	99.7173	99.5337	99.8934	99.4462

across all datasets. However, the rate of improvement diminishes beyond a certain point, indicating a saturation effect where additional samples provide marginal gains. The results highlight the tradeoff between dataset size and computational efficiency, emphasizing the importance of selecting an optimal training sample size to balance performance and resource constraints.

Fig. 3 highlights the impact of transformer layer depth on classification performance in HSIC. The results show that increasing the number of layers initially improves accuracy, but excessive depth leads to diminishing returns and potential overfitting. While deeper models can capture more complex spatial-spectral dependencies, their performance fluctuates across datasets, indicating dataset-specific optimal depths. These findings emphasize the need for a balanced architectural design to maximize accuracy while maintaining computational efficiency.

The number of attention heads directly influences the model's ability to capture spatial-spectral relationships in HSIC as shown in Table III. This test evaluates how varying attention heads affect classification accuracy. The results reveal that increasing the number of heads does not always guarantee better performance, as optimal accuracy is dataset-dependent. While some datasets benefit from higher attention diversity, others show stable or slightly fluctuating accuracy. These findings highlight the tradeoff between attention richness and model efficiency, guiding the selection of an optimal head configuration.

Table IV presents a quantitative analysis of the impact of different activation functions on the performance of *DiffFormer*. Among the tested activations, SwiGLU consistently outperforms all others, achieving the highest κ , OA, and AA across all

TABLE V
COMPARISON OF ACTIVATION FUNCTIONS FOR HSIC

Activation	Feature Selectivity	Gradient Flow	Spectral Adaptation
ReLU	No	Moderate	No
PReLU	No	Improved	No
ELU	No	Good	No
GELU	No	Best	No
SwiGLU	Yes	Best	Yes

datasets. Notably, GELU offers competitive performance, particularly on PU, but lacks the fine-grained feature sensitivity required for HSIC. Traditional activations such as ReLU, PReLU, and ELU show comparatively lower accuracy, indicating that they may not be optimal for capturing the complex spectral-spatial dependencies in HSIs. These results demonstrate that SwiGLU effectively enhances feature expressiveness, leading to superior classification accuracy.

Table V further provides a qualitative assessment of activation functions in HSIC based on three critical properties: feature selectivity, gradient flow, and spectral adaptation. Unlike ReLU, PReLU, ELU, and GELU, SwiGLU uniquely supports feature selectivity and spectral adaptation, which are crucial for hyperspectral data representation. In addition, SwiGLU exhibits the best gradient flow, similar to GELU, mitigating vanishing gradients and improving model stability. The ability of SwiGLU to selectively amplify relevant spectral-spatial features while maintaining smooth gradient propagation explains its superior classification performance in Table IV. These findings highlight the importance of activation function choice in HSIC and establish SwiGLU as the most effective activation for *DiffFormer*, facilitating robust spectral-spatial representation learning.

Overall, our analysis demonstrates that attention mechanisms, patch size selection, and training sample allocation significantly impact HSIC performance. DMHSA consistently outperforms other attention strategies by effectively capturing spectral-spatial dependencies, while optimal patch sizes strike a balance between fine-grained feature extraction and contextual information. In addition, increasing training samples improves classification accuracy, though with diminishing returns beyond a threshold. These findings underscore the importance of designing efficient attention-based models with well-calibrated input configurations to maximize classification accuracy and computational efficiency.

V. COMPARATIVE RESULTS AND DISCUSSION

This section presents a detailed discussion of the comparative results of *DiffFormer* against several SOTA methods. The selected comparative approaches encompass advanced architectures that exploit spatial-spectral information, transformer-based designs, hybrid models, and state-space models (Mamba). Specifically, we compare *DiffFormer* with the attention graph convolutional network (AGCN) [46], the pyramid hierarchical SST (PyFormer) [47], the spatial-spectral transformer with conditional position encoding (Former) [39], the spectral-spatial wavelet transformer (WaveFormer) [41], the hybrid convolution transformer (HViT) [48], the multihead spatial-spectral

TABLE VI
PERFORMANCE COMPARISON ON THE HC DATASET ACROSS CLASSWISE ACCURACIES, AGGREGATE METRICS, AND COMPUTATIONAL TIME

Class	AGCN	Former	PyFormer	WaveFormer	HViT	MHMamba	WaveMamba	<i>DiffFormer</i>
Strawberry	98.7257	99.2548	99.5230	99.7540	99.5305	94.9850	98.8524	99.6199
Cowpea	99.1210	99.0477	99.4872	99.0770	99.6484	90.0673	98.1687	99.7509
Soybean	99.0926	99.7731	99.7083	99.4815	99.8379	92.2877	97.6020	99.2546
Sorghum	98.6924	99.5018	98.8792	100	99.8754	93.8978	99.2528	99.9377
Water spinach	91.6666	98.8888	99.7222	99.7222	100	70.5555	97.7777	99.7222
Watermelon	95.2941	96.2500	98.0882	86.3970	92.3529	48.1617	85.4411	94.8529
Greens	95.4827	98.8706	99.7741	97.8543	98.5883	93.6758	98.7577	98.0237
Trees	98.7391	98.5722	99.7218	97.5523	97.1815	85.9447	95.6054	99.5735
Grass	95.6001	99.4720	99.7536	98.3104	98.5568	86.5188	97.2896	98.9792
Red roof	99.9049	99.8098	99.8415	99.6830	99.9366	97.3375	99.1759	99.4394
Gray roof	99.4677	99.4874	99.7831	98.6792	99.5466	94.1060	98.0484	99.2903
Plastic	97.0108	97.6449	97.5543	89.8550	98.3695	72.3731	98.3695	98.9130
Bare soil	96.0877	96.0877	95.8318	96.1243	95.6855	75.6124	91.9561	96.9287
Road	99.6048	94.8096	99.5689	97.2701	99.0301	91.5229	98.3477	99.4073
Bright object	97.0674	97.6539	100	98.2404	98.8269	61.5835	82.9912	98.2404
Water	99.8320	99.9248	99.9734	99.9602	99.9204	99.2528	99.7038	99.9381
κ	98.6568	98.4406	99.2007	97.3726	98.5555	84.2427	97.7480	99.8664
OA	98.8519	98.7459	98.5176	99.0108	91.0975	98.0753	99.3357	99.3137
AA	97.5868	98.9308	99.5362	98.7341	99.1547	92.3918	96.0837	99.4136
Time (s)	2347.30	3606.86	85926.77	3622.55	3832.18	52495.05	36811.62	3669.13

The higher values are in bold.

Mamba' [49], and the spatial-spectral wavelet Mamba (WaveMamba) [44]. For all methods, we followed the experimental settings outlined in their respective papers while uniformly employing a 12×12 patch size for input data across all datasets. This ensures a standardized framework for evaluating classification performance, eliminating biases arising from variations in patch size.

Table VI provides a comprehensive evaluation of various SOTA models on the HC dataset, reporting per-class accuracy, OA, AA, κ , and computational time. The results demonstrate that *DiffFormer* consistently outperforms other methods across most metrics, particularly in terms of κ and OA, achieving 99.8664% and 99.3137%, respectively. Notably, *DiffFormer* achieves high per-class accuracy for critical classes such as Strawberry (99.6199%), Cowpea (99.7509%), and Sorghum (99.9377%), exceeding the performance of the best competitors such as WaveFormer and HViT. It also demonstrates competitive accuracy for challenging classes such as Water spinach (99.7222%), whereas other models, such as MHMamba, perform significantly worse (70.5555%). In addition, *DiffFormer* maintains computational efficiency with a runtime of 3669.13 s, which is significantly lower than models such as PyFormer (85926.77 s) and MHMamba (52495.05 s). The qualitative results presented in Fig. 4 further corroborate the quantitative findings, showing that *DiffFormer* produces more accurate and spatially coherent classification maps compared to its counterparts. Models such as MHMamba and WaveMamba exhibit noticeable misclassifications and noisy outputs, particularly in heterogeneous regions. In contrast, *DiffFormer* achieves smooth and precise segmentation, effectively capturing spatial-spectral relationships.

Table VII summarizes the performance across various land cover classes in the UH dataset. Moreover, Fig. 5 illustrates the classification maps produced by each model, highlighting spatial variability and class-specific performance. The results show that the proposed *DiffFormer* achieves competitive accuracy across most classes, outperforming other models in critical categories. For example, *DiffFormer* attains a perfect accuracy (100%) for high-precision classes such as Synthetic Grass, Highway, and Running Track, demonstrating its robustness in distinguishing spectrally and spatially homogenous regions. Notably, the

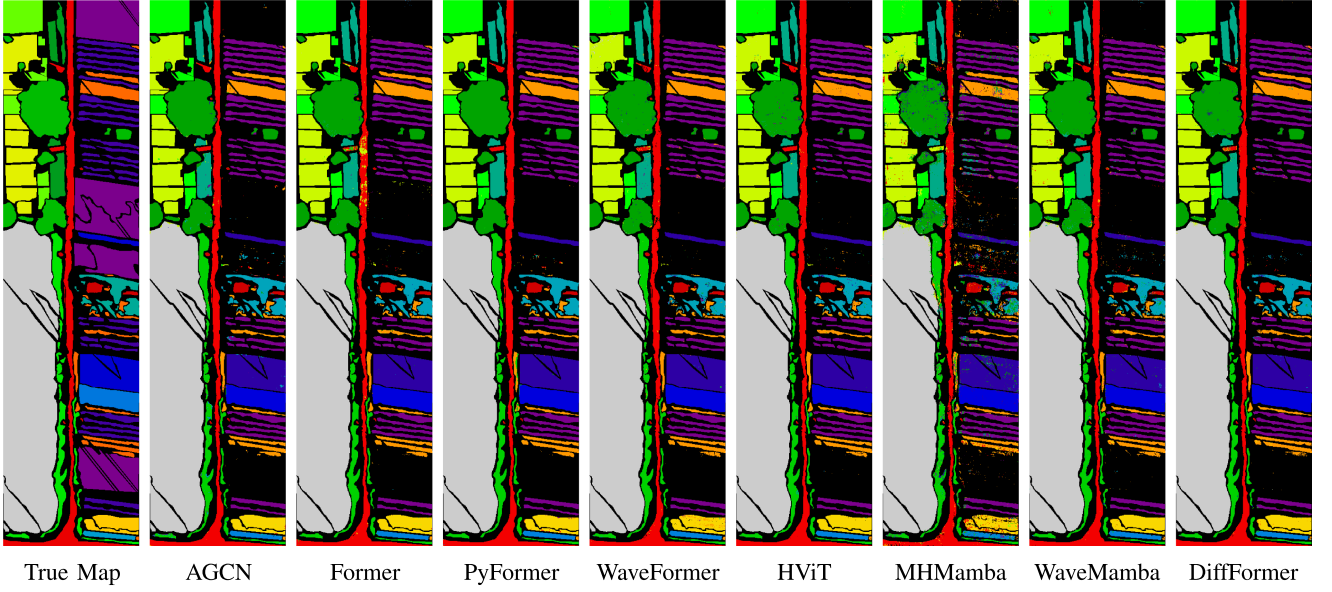


Fig. 4. Classification maps for the HC dataset, highlighting spatial variability and class-specific performance.

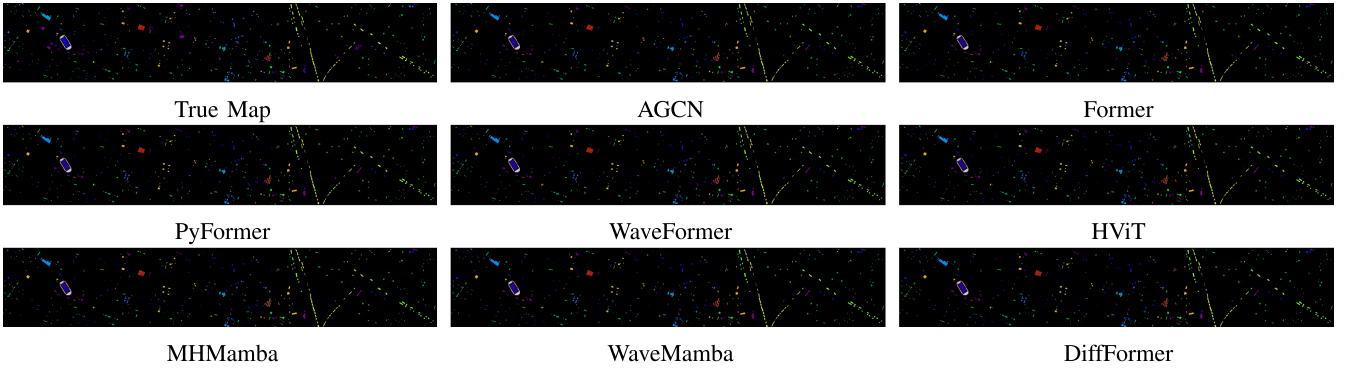


Fig. 5. Classification maps for the UH dataset, highlighting spatial variability and class-specific performance.

TABLE VII
PERFORMANCE COMPARISON ON THE UH DATASET ACROSS CLASSWISE
ACCURACIES, AGGREGATE METRICS, AND COMPUTATIONAL TIME

Class	AGCN	Former	PyFormer	WaveFormer	HViT	MHMamba	WaveMamba	DiffFormer
Healthy grass	99.3610	99.5207	100	99.5207	98.8817	99.0415	98.4025	98.9333
Stressed grass	99.8405	99.8405	99.5215	99.8405	99.8405	97.7671	99.8405	99.2021
Synthetic grass	99.7126	100	100	100	100	99.7126	99.7126	100
Trees	99.8392	100	98.8745	99.8392	99.6784	96.3022	99.6784	100
Soil	99.6779	100	99.8389	100	99.8389	99.1948	100	100
Water	98.7654	100	100	100	98.7654	96.2962	99.3827	96.9387
Residential	98.8958	97.9495	98.4227	98.4227	97.4763	84.7003	96.6876	99.2105
Commercial	95.9807	99.0353	95.0160	99.8392	99.5176	93.5691	99.3569	100
Road	99.3610	99.0415	99.8402	99.0415	99.0415	92.6517	98.4025	99.7340
Highway	100	100	100	100	100	88.4364	99.5114	100
Railway	99.6763	100	94.8220	100	99.6763	96.7637	99.8381	100
Parking Lot 1	99.5137	99.1896	98.5413	99.3517	99.1896	91.7341	99.3517	100
Parking Lot 2	99.1452	91.4529	93.1623	91.8803	94.4444	62.8205	94.4444	97.8723
Tennis Court	100	99.5327	100	99.5327	100	96.7289	99.0654	100
Running Track	100	100	100	100	100	100	100	100
κ	99.2231	99.2086	98.4774	99.3381	99.1655	93.1787	98.9498	99.5923
OA	99.2314	99.2681	98.5362	99.3878	99.2282	93.6926	99.0286	99.6239
AA	99.3179	99.0375	98.5359	99.1512	99.0900	93.0479	98.9116	99.4594
Time (s)	98.86	6392.95	1001.73	146.79	152.05	362.30	1374.47	219.45

model also achieves superior performance in challenging categories such as Commercial (100%) and Residential (99.2105%), where other models show variability. In contrast, models such as MHMamba underperform in several categories, such as Parking Lot 2 (62.8205%) and Water (96.2962%), indicating its limitations in handling complex spectral-spatial variations.

Similarly, HViT exhibits consistent but slightly lower performance across most classes, particularly for Road (99.0415%) and Tennis Court (100%). From an aggregate perspective, *DiffFormer* outperforms all competing models, achieving the highest OA (99.6229%), AA (99.4594%), and κ (99.5923%). The second-best model, WaveFormer, delivers comparable performance but falls slightly short in terms of AA (99.1512%) and OA (99.3878%), underscoring the improvements introduced by *DiffFormer*. In terms of computational efficiency, *DiffFormer* exhibits a balanced tradeoff between accuracy and time, with a processing time of 219.45 s. While WaveFormer is marginally faster (146.79 s), its slightly lower performance highlights the significance of the proposed enhancements in *DiffFormer*. Notably, Former requires the longest processing time (6392.95 s), making it impractical for large-scale applications.

The results in Table VIII comprehensively evaluate and compare the performance of various models on the SA dataset. From Table VIII, the proposed model, *DiffFormer*, outperforms other models in most aggregate metrics, achieving the highest OA (99.8152%), AA (99.8546%), and κ (99.7942%). These results

TABLE VIII
PERFORMANCE COMPARISON ON THE SA DATASET ACROSS CLASS-WISE
ACCURACIES, AGGREGATE METRICS, AND COMPUTATIONAL TIME

Class	AGCN	Former	PyFormer	WaveFormer	HViT	MHMamba	WaveMamba	DiffFormer
Broccoli 1	100	100	100	100	100	99.8007	100	100
Broccoli 2	100	100	100	100	100	100	100	100
Fallow	100	100	100	100	100	99.2914	99.8987	100
Fallow Rough	99.7130	99.8565	100	99.7130	99.8565	99.5695	100	99.5215
Fallow Smooth	100	99.9253	100	99.9253	99.7012	98.0582	99.1038	99.0037
Stubble	100	100	100	100	100	100	99.9494	100
Celery	99.9441	100	100	100	100	99.7206	99.9441	100
Grapes	99.4854	99.0951	99.5919	99.1838	99.0596	97.1965	99.2193	99.8521
Soil Vineyard	99.9677	100	100	100	100	100	100	99.8925
Corn Senesced	99.7559	99.9389	99.8779	99.8779	99.8779	99.3288	99.9389	100
Lettuce 4wk	99.4382	100	98.8764	100	100	99.6254	99.8127	100
Lettuce 5wk	99.8961	100	100	100	100	100	100	100
Lettuce 6wk	100	100	100	100	100	99.3449	100	100
Lettuce 7wk	99.4392	99.6261	99.6261	100	100	98.5046	99.8130	100
Vineyard Untrained	99.4496	98.4865	90.6989	98.2388	98.2663	73.6378	98.4865	99.4039
Vineyard Vertical	98.7818	100	99.7785	100	100	99.3355	100	100
κ	99.6914	99.5433	98.4639	99.5277	99.4815	95.9601	99.5186	99.7942
OA	99.7228	99.5898	98.6218	99.5714	99.5344	95.6068	99.5677	99.8152
AA	99.7419	99.8080	99.2781	99.8021	99.7861	97.7134	99.7604	99.8546
Time (s)	328.86	561.87	3590.99	529.59	533.54	1477.32	8526.14	767.71

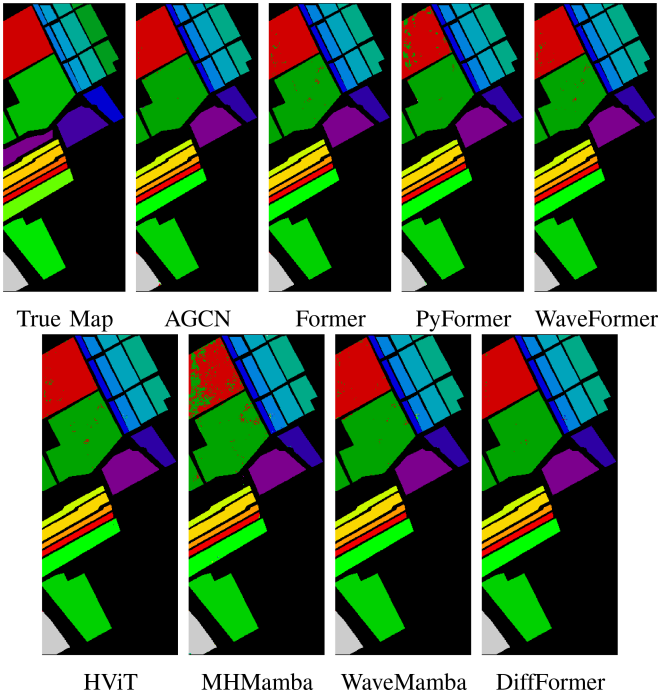


Fig. 6. Classification maps for the SA dataset, highlighting spatial variability and class-specific performance.

indicate superior classification reliability. Classwise accuracies reveal that *DiffFormer* maintains consistent performance across classes, with perfect scores for challenging categories such as “Broccoli 1,” “Broccoli 2,” and “Stubble,” outperforming models such as MHMamba, which struggles particularly in “Vineyard Untrained” with a significantly lower accuracy (73.6378%). Furthermore, *DiffFormer* demonstrates competitive computational efficiency with a runtime of 767.71 s, which, while not the fastest, remains practical compared to WaveMamba’s extensive runtime of 8526.14 s.

Fig. 6 provides a qualitative comparison of classification maps. It highlights the spatial consistency and accuracy of *DiffFormer*, especially in regions with complex class distributions, such as “Fallow Rough” and “Vineyard Vertical.” Misclassifications, evident in other models such as PyFormer and MHMamba, are significantly reduced in *DiffFormer*, leading to smoother and more accurate spatial patterns. In addition, the high fidelity of the maps corroborates the quantitative superiority of *DiffFormer*.

TABLE IX
PERFORMANCE COMPARISON ON THE PU DATASET ACROSS CLASS-WISE
ACCURACIES, AGGREGATE METRICS, AND COMPUTATIONAL TIME

Class	AGCN	Former	PyFormer	WaveFormer	HViT	MHMamba	WaveMamba	DiffFormer
Asphalt	100	99.0349	99.8190	98.8540	98.6429	92.8226	98.9143	99.4972
Meadows	99.9678	99.9249	99.9678	99.9249	99.9356	98.6593	100	99.9642
Gravel	98.2840	93.6129	97.7121	93.8036	92.6596	83.9847	95.2335	96.5079
Trees	99.3472	99.5430	99.6736	99.2819	99.2167	92.8198	99.2167	99.2383
Metal Sheets	100	100	99.8514	100	100	99.7028	100	100
Bare Soil	99.9204	99.9602	99.6022	99.7613	99.7215	81.7422	100	99.7349
Bitumen	99.8496	99.2481	99.5488	99.0977	99.0977	81.8045	98.7969	99.7493
Bricks	99.4024	96.4149	95.7088	96.1434	95.5458	74.7963	96.4149	98.0090
Shadows	98.7341	98.9451	100	98.3122	97.8902	94.5147	98.5232	99.6478
κ	99.6839	98.8229	99.0982	98.3532	98.0789	89.2972	98.8910	99.2872
OA	99.7615	99.1116	99.1882	98.6865	98.4816	91.9908	99.1630	99.4623
AA	99.5006	98.5204	99.3875	99.0087	98.8544	88.9830	98.5666	99.1498
Time (s)	261.21	383.60	2840.00	383.09	405.36	1041.99	3952.37	601.86

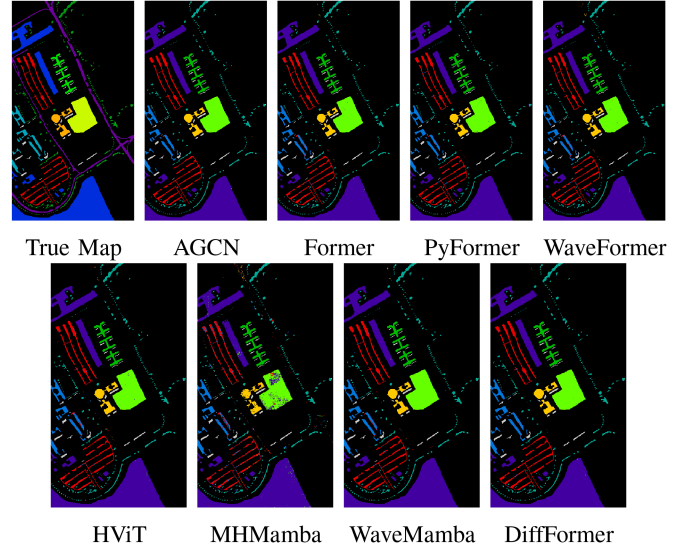


Fig. 7. Classification maps for the PU dataset, highlighting spatial variability and class-specific performance.

Table IX presents a comprehensive comparison of classification performance on the PU dataset. Notably, *DiffFormer* demonstrates superior OA of 99.4623%, κ of 99.2872%, and AA of 99.1498%, consistently outperforming existing approaches. In particular, *DiffFormer* achieves SOTA results for challenging classes such as Asphalt (99.4972%), Bitumen (99.7493%), and Shadows (99.6478%). These improvements suggest its effectiveness in capturing spatial and spectral correlations. In contrast, AGCN also achieves high performance, with the highest AA of 99.5006%, but marginally lower OA and κ . The Former, PyFormer, and WaveFormer exhibit competitive results across most metrics but fall behind *DiffFormer*, especially in the Bare Soil and Bitumen classes. WaveMamba and MHMamba, while innovative, exhibit suboptimal performance, particularly in classes such as Bricks and Gravel, indicating potential limitations in capturing finer spatial features. From a computational perspective, *DiffFormer* balances accuracy with efficiency, achieving competitive inference time (601.86 s) relative to Former (383.60 s) and WaveFormer (383.09 s), while significantly outperforming models such as PyFormer (2840.00 s) and WaveMamba (3952.37 s).

The classification maps in Fig. 7 visually highlight the spatial distribution of predicted labels for each model. *DiffFormer* showcases the superior performance of class boundaries and

reduction in misclassified pixels, especially in heterogeneous regions such as Meadows and Bare Soil. In comparison, AGCN and WaveFormer exhibit satisfactory classification but suffer from subtle boundary inconsistencies. Models such as MH-Mamba and WaveMamba display noticeable artifacts and reduced precision in high-variability regions, consistent with their lower quantitative performance. Conversely, Former and PyFormer demonstrate balanced accuracy across large homogeneous classes but struggle to preserve spatial details in smaller regions such as Bitumen and Bricks. The visualization further validates the effectiveness of *DiffFormer* in handling spectral-spatial variability, ensuring high classification accuracy across diverse land cover types.

VI. CONCLUSION

This article introduced *DiffFormer*, a novel differential SST designed to address the challenges of spectral redundancy and spatial discontinuity in HSIC. The core innovation of *DiffFormer* lies in its DMHSA mechanism, which enhances hyperspectral feature representation by refining spectral-spatial dependencies through differential attention scoring. In addition, the integration of the SwiGLU activation function and class token-based aggregation improves feature discrimination while maintaining computational efficiency. The proposed patch-based spectral-spatial tokenization strategy ensures scalable feature extraction by optimizing input dimensionality without compromising spectral fidelity. Extensive experiments on benchmark hyperspectral datasets validate the superiority of *DiffFormer* over SOTA models in both classification accuracy and computational efficiency. Empirical results demonstrate that *DiffFormer* effectively balances spectral feature expressivity and model complexity, achieving competitive performance with a reduced computational footprint. The ablation studies further highlight the impact of DMHSA, SwiGLU, and spectral-spatial tokenization in improving generalization and robustness across different HSI scenarios.

Despite these advancements, several future research directions remain open. One promising avenue is the development of adaptive or hierarchical tokenization strategies that dynamically adjust the granularity of spectral-spatial feature extraction based on scene complexity, reducing redundant computations while preserving fine-grained spectral details. In addition, optimizing transformer architectures for energy efficiency and hardware acceleration could enable real-time HSIC on resource-constrained platforms such as UAVs, CubeSats, and edge devices. Another important direction is the integration of self-supervised or contrastive learning paradigms to enhance feature robustness, particularly in scenarios where labeled hyperspectral data is scarce. Leveraging spectral similarity constraints or contrastive embeddings could improve generalization while reducing dependency on extensive manual annotations. Furthermore, exploring graph-based or attention fusion techniques could enhance spatial-spectral reasoning by explicitly modeling interpixel relationships, improving classification accuracy in complex land cover or remote sensing applications. Finally, incorporating uncertainty quantification

techniques within *DiffFormer* could provide better interpretability and reliability in decision-making, ensuring its applicability to critical remote sensing tasks such as disaster monitoring, precision agriculture, and mineral exploration.

REFERENCES

- [1] S. Faisal, M. P.-L. Ooi, S. K. Abeyssekera, Y.-C. Kuang, and D. Fletcher, "Roadmap for measurement and applications: Uncertainty quantification and visualization for optimal decision-making in hyperspectral imaging-based precision agriculture," *IEEE Instrum. Meas. Mag.*, vol. 28, no. 1, pp. 23–32, Feb. 2025.
- [2] Y. Zhang, L. Liu, and X. Yang, "Hyperspectral image classification using spectral-spatial dual random fields with Gaussian and Markov processes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 4199–4212, 2025.
- [3] R. Rajabi, A. Zehtabian, K. D. Singh, A. Tabatabaenejad, P. Ghamisi, and S. Homayouni, "Hyperspectral imaging in environmental monitoring and analysis," *Front. Environ. Sci.*, vol. 11, 2024, Art. no. 1353447.
- [4] J. Chen, J. Li, and P. Gamba, "Adaptive multi-task autoencoder-based hyperspectral unmixing exploiting auxiliary data via graph associations," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5509213.
- [5] Y. Feng, X. Yi, S. Wang, J. Yue, S. Xia, and L. Fang, "HyperEDL: Spectral-spatial evidence deep learning for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5508316.
- [6] M. H. Khan et al., "Hyperspectral imaging-based unsupervised adulterated red chili content transformation for classification: Identification of red chili adulterants," *Neural Comput. Appl.*, vol. 33, no. 21, pp. 14507–14521, 2021.
- [7] M. H. Khan, Z. Saleem, M. Ahmad, A. Sohaib, H. Ayaz, and M. Mazzara, "Hyperspectral imaging for color adulteration detection in red chili," *Appl. Sci.*, vol. 10, no. 17, 2020, Art. no. 5955.
- [8] Z. Saleem, M. H. Khan, M. Ahmad, A. Sohaib, H. Ayaz, and M. Mazzara, "Prediction of microbial spoilage and shelf-life of bakery products through hyperspectral imaging," *IEEE Access*, vol. 8, pp. 176986–176996, 2020.
- [9] M. H. F. Butt, H. Ayaz, M. Ahmad, J. P. Li, and R. Kuleev, "A fast and compact hybrid CNN for hyperspectral imaging-based bloodstain classification," in *Proc. 2022 IEEE Congr. Evol. Comput.* IEEE, 2022, pp. 1–8.
- [10] M. Zulfikar, M. Ahmad, A. Sohaib, M. Mazzara, and S. Distefano, "Hyperspectral imaging for bloodstain identification," *Sensors*, vol. 21, no. 9, 2021, Art. no. 3045.
- [11] H. Ayaz, M. Ahmad, M. Mazzara, and A. Sohaib, "Hyperspectral imaging for minced meat classification using nonlinear deep features," *Appl. Sci.*, vol. 10, no. 21, 2020, Art. no. 7783.
- [12] H. Ayaz et al., "Myoglobin-based classification of minced meat using hyperspectral imaging," *Appl. Sci.*, vol. 10, no. 19, 2020, Art. no. 6862.
- [13] M. Ahmad et al., "Hyperspectral image classification—Traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.
- [14] P. Addabbo, N. Fiscante, G. Giunta, D. Orlando, G. Ricci, and S. L. Ullo, "Multiple sub-pixel target detection for hyperspectral imaging systems," *IEEE Trans. Signal Process.*, vol. 71, pp. 1599–1611, 2023.
- [15] M. Ahmad et al., "Spatial-spectral morphological mamba for hyperspectral image classification," *Neurocomputing*, vol. 636, 2025, Art. no. 129995. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231225006678>
- [16] J. Feng et al., "S4dI: Shift-sensitive spatial-spectral disentangling learning for hyperspectral image unsupervised domain adaptation," 2024, *arXiv:2408.15263*.
- [17] Y. Ding et al., "Slcgc: A lightweight self-supervised low-pass contrastive graph clustering network for hyperspectral images," 2025, *arXiv:2502.03497*.
- [18] Y. Ding et al., "Self-supervised locality preserving low-pass graph convolutional embedding for large-scale hyperspectral image clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536016.
- [19] Y. Ding et al., "Adaptive homophily clustering: Structure homophily graph learning with adaptive filter for hyperspectral image," 2025, *arXiv:2501.01595*.
- [20] Z. Zhao, X. Xu, S. Li, and A. Plaza, "Hyperspectral image classification using groupwise separable convolutional vision transformer network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5511817.

- [21] S. Cheng, R. Chan, and A. Du, "MS2I2Former: Multiscale spatial-spectral information interactive transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5532919.
- [22] F. Xu, S. Mei, G. Zhang, N. Wang, and Q. Du, "Bridging CNN and transformer with cross-attention fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5522214.
- [23] S. Jia, Y. Wang, S. Jiang, and R. He, "A center-masked transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5510416.
- [24] T. Arshad and J. Zhang, "Hierarchical attention transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5504605.
- [25] M. Jiang et al., "GraphGST: Graph generative structure-aware transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5504016.
- [26] L. Huang, Y. Chen, and X. He, "Foundation model-based spectral-spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5529825.
- [27] Z. Shu, Y. Wang, and Z. Yu, "Dual attention transformer network for hyperspectral image classification," *Eng. Appl. Artif. Intell.*, vol. 127, 2024, Art. no. 107351. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S095219762301535X>
- [28] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514715.
- [29] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "QTN: Quaternion transformer network for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7370–7384, Dec. 2023.
- [30] 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, 2023, Art. no. 5517617.
- [31] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715.
- [32] 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, 2022, Art. no. 5532513.
- [33] B. Zhang, Y. Chen, Y. Rong, S. Xiong, and X. Lu, "MATNet: A combining multi-attention and transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5506015.
- [34] T. Ye et al., "Differential transformer," 2024, *arXiv:2410.05258*.
- [35] M. Ye, J. Chen, F. Xiong, and Y. Qian, "Adaptive graph modeling with self-training for heterogeneous cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5503815.
- [36] M. Ahmad, M. Usama, M. Mazzara, S. Distefano, H. A. Altuwaijri, and S. L. Ullo, "Fusing transformers in a tuning fork structure for hyperspectral image classification across disjoint samples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 18167–18181, 2024.
- [37] J. Fang, J. Yang, A. Khader, and L. Xiao, "MIMO-SST: Multi-input multi-output spatial-spectral transformer for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5510020.
- [38] W. Huang, Y. Deng, S. Hui, Y. Wu, S. Zhou, and J. Wang, "Sparse self-attention transformer for image inpainting," *Pattern Recognit.*, vol. 145, 2024, Art. no. 109897.
- [39] M. Ahmad, M. Usama, A. M. Khan, S. Distefano, H. A. Altuwaijri, and M. Mazzara, "Spatial-spectral transformer with conditional position encoding for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5508205.
- [40] Y. Sun, X. Zhi, S. Jiang, G. Fan, X. Yan, and W. Zhang, "Image fusion for the novelty rotating synthetic aperture system based on vision transformer," *Inf. Fusion*, vol. 104, 2024, Art. no. 102163.
- [41] M. Ahmad, U. Ghous, M. Usama, and M. Mazzara, "WaveFormer: Spectral-spatial wavelet transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5502405.
- [42] T. Kim, J. Kim, H. Oh, and J. Kang, "Deep transformer based video inpainting using fast Fourier tokenization," *IEEE Access*, vol. 12, pp. 21723–21736, 2024.
- [43] Y. Shi, J. Xia, M. Zhou, and Z. Cao, "A dual-feature-based adaptive shared transformer network for image captioning," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 5009613.
- [44] M. Ahmad, M. Usama, M. Mazzara, and S. Distefano, "WaveMamba: Spatial-spectral wavelet mamba for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, 2025, Art. no. 5500505.
- [45] Y. Zhong et al., "Mini-UAV-borne hyperspectral remote sensing: From observation and processing to applications," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 4, pp. 46–62, Dec. 2018.
- [46] A. Jamali, S. K. Roy, D. Hong, P. M. Atkinson, and P. Ghamisi, "Attention graph convolutional network for disjoint hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5503005.
- [47] 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.
- [48] J. Z. T. Arshad and I. Ullah, "A hybrid convolution transformer for hyperspectral image classification," *Eur. J. Remote Sens.*, vol. 57, no. 1, 2024, Art. no. 2330979.
- [49] M. Ahmad et al., "Multi-head spatial-spectral Mamba for hyperspectral image classification," *Remote Sens. Lett.*, vol. 16, no. 4, pp. 15–29, 2025, doi: [10.1080/2150704X.2025.2461330](https://doi.org/10.1080/2150704X.2025.2461330).