DSFMNet: A Dual-Scale Frequency-Modulated Network With a Linear-Complexity Hybrid Architecture for Hyperspectral Image Classification

Junjie Li[1], Ziping He*[1], Ningjuan Wang[1], Baokai Zu[2], Hongguang Xiao[1], Yuchuan Chen[1]

[1]*School of Computer Science and Technology, Changsha University of Science and Technology, Changsha 410114, China*

[2]*College of Computer Science, Beijing University of Technology, Beijing 100124, China*

{*202308010307, heziping, 003331*}*@csust.edu.cn; wangningjuan38@gmail.com; bzu@bjut.edu.cn; chuan990621@163.com*

***Abstract*—Hyperspectral image (HSI) classification demands efficient modeling of spatial–spectral–frequency information, but existing methods often struggle to balance accuracy and computational cost. In this paper, we propose DSFMNet, a dual-scale frequency-modulated linear-complexity network that comprises a fine-scale branch and a coarse-scale branch. Specifically, for the fine-scale branch, we introduce the Cross-Flow Frequency-Modulated Vision Mamba (CF-FMViM) module, which integrates state space duality (SSD)-based hidden state mixing with local frequency modulation to preserve detailed spatial–spectral structures. Complementarily, for the coarse-scale branch, we design the Frequency-Modulated Token-Selective Statistical Attention (FM-TSSA) module, which combines token-selective statistical attention with frequency modulation to enable efficient global semantic modeling, reducing the complexity from quadratic to linear in the number of tokens. The fusion of these two branches enables synergistic optimization across spatial, spectral, and frequency domains. Extensive experiments on the Houston2013, Whu-Hi-HongHu, and Houston2018 datasets demonstrate that DSFMNet consistently outperforms state-of-the-art methods while significantly reducing computational costs.**

***Index Terms*—Hyperspectral image classification, Transformer Networks, linear attention, frequency modulation.**

I. Introduction

Hyperspectral images (HSI) provide rich cross-spectral information crucial for distinguishing fine-grained materials. Their utility is well-established in remote sensing applications such as mineral exploration [1, 2], urban planning [3, 4], and environmental assessment [5, 6]. Leveraging spectral information across hundreds of continuous bands, significant progress has been made in HSI classification [7–10].

Recent deep learning approaches [10–14] have shown encouraging results in this domain. Specifically, CNN-based methods [15, 16] utilize local receptive fields to effectively extract texture and detailed features. While the inherent locality of CNNs limits their capacity to model long-range dependencies in complex land cover scenarios, Vision Transformers (ViTs) [17] effectively address this via global attention. In HSI classification, Hong *et al.* [18] pioneered ViT applications, and Mei *et al.* [19] proposed a group-aware hierarchical Transformer, validating the efficacy of attention mechanisms. However, the quadratic computational cost $\mathcal{O}(n^2)$ of standard self-attention (where n is the token count) bottlenecks efficiency on high-dimensional spectral data. To overcome this, Gu and Dao [20] introduced Mamba as a linear-complexity alternative, which Zhou *et al.* [21] subsequently leveraged for efficient long-range HSI modeling.

Although the aforementioned deep learning models possess strong capabilities in feature extraction and have demonstrated significant performance in HSI classification, several critical challenges remain:

1) Representation Bottleneck in High-Dimensional HSI: The continuous redundancy of HSI bands challenges single-paradigm architectures, which struggle to distill essential semantics from high-dimensional spectral volumes, often yielding superficial features or excessive memory costs.

2) Neglect of Frequency-Domain Cues: Despite Fourier analysis [22] revealing distinct frequency characteristics, standard attention mechanisms predominantly focus on spatial processing, overlooking critical frequency information.

Recognizing that effective HSI classification demands frequency-aware modeling of both fine details and global context under strict computational budgets, we propose DSFMNet: a dual-scale frequency-modulated network with a linear-complexity hybrid architecture. Our main contributions are summarized as follows:

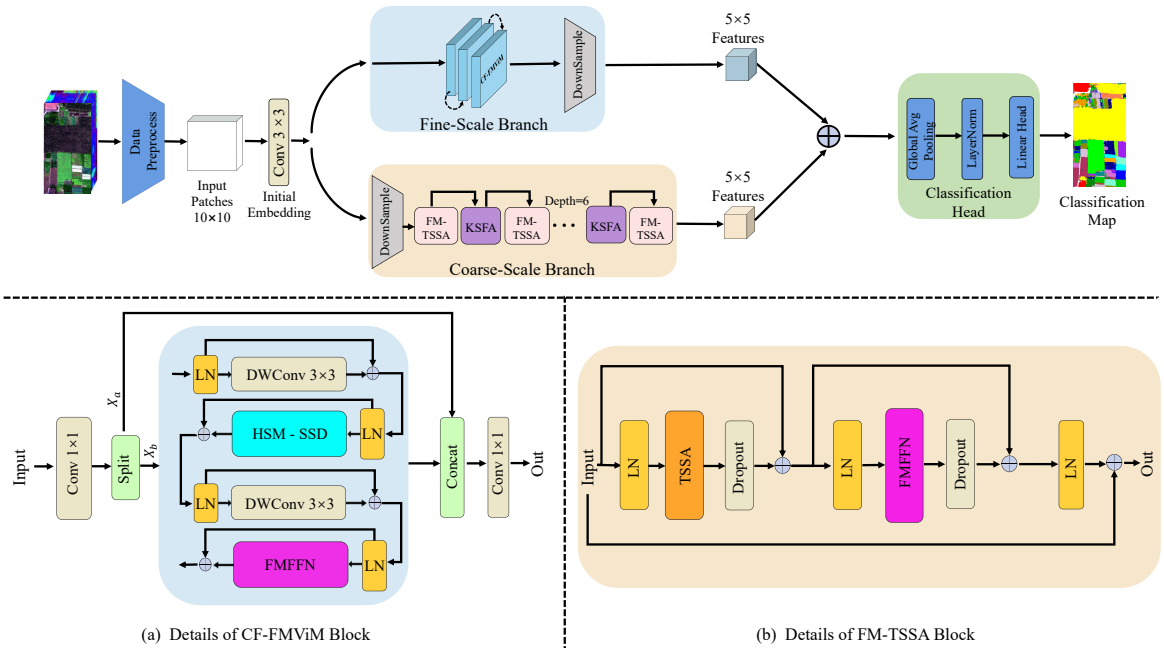
1) DSFMNet, a dual-scale frequency-modulated network with a linear-complexity hybrid architecture, is introduced. It effectively resolves the trade-off between representation capability and computational efficiency by synergistically optimizing fine-grained details and global semantic context.

2) The Cross-Flow Frequency-Modulated Vision Mamba (CF-FMViM) is designed to capture and refine fine-grained local details by extracting complementary spatial–frequency features with linear complexity, thereby preserving textures and edges under limited computational budgets.

3) The Frequency-Modulated Token-Selective Statistical Attention (FM-TSSA) is designed to model global semantic context by capturing long-range dependencies via frequency-aware linear attention, enabling robust land-cover discrimination without quadratic complexity.

4) Extensive experiments on the Houston2013 [23], Whu-Hi-HongHu [24], and Houston2018 [25] datasets demonstrate

* Corresponding author: Ziping He (heziping@csust.edu.cn).



(a) Details of CF-FMViM Block (b) Details of FM-TSSA Block

Fig. 1. The illustration of the DSFMNet. The fine-scale branch employs 3 consecutive CF-FMViM modules with residual paths; the coarse-scale branch adopts 6-depth FM-TSSA modules (also integrated with residual paths), where KSFA is used after the first and fourth modules for multi-scale information fusion. Finally, the residuals of the two branches are summed.

that DSFMNet outperforms state-of-the-art methods.

The code will be made publicly available at this link: https://github.com/ljj5525/IJCNN2026DSFMNet.

II. METHOD

A. Overall Architecture

Given an original hyperspectral image $\mathbf{X} \in \mathbb{R}^{\mathcal{C}\times\mathcal{H}\times\mathcal{W}}$, we first reduce the redundant spectral dimension to C_{in} via PCA and extract overlapping patches $P \in \mathbb{R}^{C_{in}\times H\times W}$. A 3×3 convolution then projects each patch into a unified initial embedding $X_0 \in \mathbb{R}^{C\times H\times W}$, where the hidden dimension is set to $C = 64$.

As illustrated in Fig. 1, DSFMNet processes X_0 through a dual-granularity architecture that decouples detail-preserving texture modeling from efficient semantic abstraction. The fine-grained branch preserves local textures while expanding receptive fields via cross-stage feature reuse and residual refinement. Each stage employs CF-FMViM, where an input-adaptive state-space mixer (HSM-SSD) enables long-range dependency propagation, and an FMFFN enhances spectral selectivity [26, 27]. Multi-scale alignment is achieved through subsequent downsampling. Concurrently, the coarse-grained branch downsamples X_0 and applies FM-TSSA blocks [28]—which synergize statistical linear attention with frequency modulation—to capture linear-complexity global semantics robust to spectral variability. To offset the reduced spatial inductive bias of linear attention, KSFA blocks [29] are sparsely injected at two abstraction levels. Ultimately, both branch outputs are fused via residual addition and passed to a classification head for prediction.

B. CF-FMViM: Cross-Flow Frequency-Modulated Vision Mamba

The fine-grained branch employs our proposed CF-FMViM module, which integrates three key innovations: A Cross-Flow (CF) architecture for cross-stage feature propagation; HSM-SSD for efficient state space modeling; FMFFN for frequency-domain enhancement.

1) Cross-Flow (CF) Architecture: The CF structure employs a split-iterate-merge mechanism for efficient cross-scale propagation. An input $X_0 \in \mathbb{R}^{C\times H\times W}$ undergoes 1×1 convolutional expansion and splits into a cross-flow path X_a and a processing path X_b. Path X_b sequentially traverses n FMViM sub-blocks (instantiated with $n = 1$ in our work) to yield outputs $Y_1, \ldots, Y_n$. These are aggregated as $X_{\text{out}} = \text{Concat}([X_a, Y_1, \ldots, Y_n])$ and refined by a final 1×1 projection. This design effectively enhances multi-stage gradient flow and feature reuse with minimal overhead.

2) Module Structure: Each FMViM block processes the input $X_0 = X_b$ through a sequence of four operations, linked by layer-wise learnable gated residuals. For $i \in \{1, 2, 3, 4\}$, the update rule is defined as:

$$X_i = \big(1 - \sigma(\boldsymbol{\alpha}_i)\big) \odot X_{i-1} \; + \; \sigma(\boldsymbol{\alpha}_i) \odot \mathcal{F}_i(X_{i-1})$$

where $\boldsymbol{\alpha}_i$ denotes the learnable scaling factors for the i-th step. The specific operations $\mathcal{F}_1$ to $\mathcal{F}_4$ correspond sequentially to $\text{DWConv}_{3\times3}(\cdot)$, HSM-SSD(LayerNorm$(\cdot)$), $\text{DWConv}_{3\times3}(\cdot)$, and FMFFN(LayerNorm$(\cdot)$). The final output of the block is X_4.

3) HSM-SSD: We propose the Hidden State Mixer with State Space Duality (HSM-SSD) to capture global spectral-spatial

dependencies while preserving local textures. Unlike standard 1D causal SSMs, HSM-SSD leverages input-dependent modulation for adaptive state transitions and global aggregation. Given the flattened input $\mathbf{X}_1^{\text{LN}} \in \mathbb{R}^{D_0 \times L_0}$ ($L_0 = H \times W$), sequential linear and 2D depthwise convolutional (DWConv_{2D}) layers generate a parameter tensor $\mathbf{P} \in \mathbb{R}^{3N \times L_0}$. Setting the state dimension $N = 64$, $\mathbf{P}$ is partitioned into input modulation $\mathbf{B}$, output projection $\mathbf{C_{mat}}$, and transition bias $\mathcal{T}$:

$$\mathbf{B}, \mathbf{C_{mat}}, \mathcal{T} = \text{Split}\left(\text{DWConv}_{2D}(\text{Linear}(\mathbf{X}_1^{\text{LN}}))\right) \quad (1)$$

We compute normalized weights $\mathbf{A} = \text{Softmax}(\mathcal{T}, \text{dim} = 0)$ to modulate $\mathbf{B}$, yielding $\mathbf{\Phi} = \mathbf{A} \odot \mathbf{B} \in \mathbb{R}^{N \times L_0}$. The input features are then mapped into the latent state space using $\mathbf{\Phi}^\top$:

$$\mathbf{H} = \mathbf{X}_1^{\text{LN}} \mathbf{\Phi}^\top \quad \in \mathbb{R}^{D_0 \times N} \quad (2)$$

This operation effectively aggregates spatial information from all L_0 tokens into N compact state prototypes. To enhance representation, $\mathbf{H}$ is projected to an expanded dimension $D_{\text{inner}} = E \times D_0$ to generate $[\mathbf{H}', \mathbf{Z}]$, followed by a residual Gated Linear Unit (GLU) scaled by a learnable parameter s:

$$\tilde{\mathbf{H}} = (\mathbf{H}' \odot \text{SiLU}(\mathbf{Z})) + s \cdot \mathbf{H}' \quad (3)$$

Finally, the output $\mathbf{Y}$ is reconstructed by linearly projecting the hidden state $\tilde{\mathbf{H}}$ back to the target dimension D and subsequently multiplying it by the output matrix $\mathbf{C_{mat}}$.

4) FMFFN: We propose the Frequency-Modulated Feed-Forward Network (FMFFN) to enhance local frequency representation. The input $\mathbf{X}_3^{\text{LN}} \in \mathbb{R}^{D_0 \times H \times W}$ is partitioned into $S \times S$ ($S = 5$) windows and processed via a 2D real-valued fast Fourier transform (RFFT). After adaptively adjusting the spectral amplitude and phase using element-wise learnable complex weights $\mathbf{W}_{\text{complex}}$, an inverse RFFT (IRFFT) yields the spatial features $\mathbf{X}_{\text{spatial}}$. These are subsequently fused via a feed-forward block:

$$\mathbf{Y}' = \text{Conv}_{1\times1}(\text{GELU}\big(\text{Conv}_{1\times1}(\text{Merge}(\mathbf{X}_{\text{spatial}}))\big)). \quad (4)$$

This mechanism enables the adaptive extraction of discriminative features based on regional frequency patterns, such as distinguishing texture-rich vegetation from smooth surfaces.

C. FM-TSSA: Frequency-Modulated Token-Selective Statistical Attention

To mitigate the computational overhead of standard self-attention, we propose the FM-TSSA module, which integrates Token-Selective Statistical Attention (TSSA) [28] for linear-complexity global modeling and a Frequency-Modulated Feed-Forward Network (FMFFN) [27] for spectral analysis. Following a Pre-Norm residual topology, each sub-layer incorporates Layer Normalization, dropout ($p = 0.1$), and skip connections to effectively balance efficient global context capture with training stability.

1) TSSA: TSSA reduces attention complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ by deriving weights from global statistics. Given the input $\mathbf{X}_{l-1}^{\text{LN}} \in \mathbb{R}^{D_1 \times L'}$ (L' corresponding to flattened windows), we first linearly project it to the latent space $\mathbf{W}$ and subsequently apply token-wise L_2 normalization to obtain $\hat{\mathbf{W}}$.

Unlike standard attention, we compute raw saliency scores $\mathbf{S}'$ based on the energy of normalized features summed across channels D_h:

$$\mathbf{S}' = \sum_{c=1}^{D_h} (\hat{W}_{h,c,l})^2 \quad \in \mathbb{R}^{H' \times L'} \quad (5)$$

Incorporating a learnable temperature $\tau \in \mathbb{R}^{H' \times 1}$, we first apply a head-wise Softmax to encourage inter-head competition:

$$\mathbf{\Pi}_{\text{raw}} = \text{Softmax}(\mathbf{S}' \odot \tau, \text{dim} = H') \quad (6)$$

Subsequently, we perform token-wise normalization to ensure a valid token-wise probability distribution, yielding the final attention map $\mathbf{\Pi}$:

$$\mathbf{\Pi} = \frac{\mathbf{\Pi}_{\text{raw}}}{\sum_{l=1}^{L'} \mathbf{\Pi}_{\text{raw},l} + \epsilon} \quad (7)$$

A global context vector $\mathbf{G}$ is then obtained by aggregating the squared raw features $\mathbf{W}^2$ weighted by $\mathbf{\Pi}$:

$$\mathbf{G} = \mathbf{\Pi}(\mathbf{W}^2)^\top \quad \in \mathbb{R}^{H' \times D_h} \quad (8)$$

To modulate features, we introduce a non-linear gate $\mathbf{A}_{\text{gate}} = (1 + \mathbf{G})^{-1}$. Finally, the output is generated via a negative feedback connection:

$$\mathbf{Y} = -(\mathbf{W} \odot \mathbf{\Pi}) \odot \mathbf{A}_{\text{gate}} \quad (9)$$

the resulting $\mathbf{Y}$ undergoes a final linear projection. The computational cost is dominated by the aggregation $\mathcal{O}(D_1 \cdot L')$, ensuring high efficiency.

In the feed-forward phase, the FMFFN (detailed in Section II-C) is integrated to establish a global-local synergy with the TSSA. While TSSA aggregates long-range semantics via global statistics, FMFFN refines local frequency-domain textures. This complementarity ensures the simultaneous extraction of macroscopic structures and microscopic patterns crucial for parsing complex scenes.

D. Asymmetric Dual-Scale Architecture

Leveraging the efficiency of CF-FMViM and FM-TSSA, we propose an Asymmetric Dual-Scale Architecture comprising parallel fine- and coarse-grained branches. The fine-grained branch ($L_{\text{fine}} = 3$) utilizes residual CF-FMViM blocks to preserve high-resolution details. Conversely, the coarse-grained branch stacks 6 FM-TSSA modules on downsampled features for deep semantic modeling, incorporating KSFA blocks at layers 1 and 4 to mitigate spatial loss. Finally, the branches are fused via residual addition to integrate orthogonal representations.

Notably, by strictly adhering to linear-complexity paradigms in both paths, our design effectively dispels the common assumption that dual-branch architectures inherently incur prohibitive computational overhead.

TABLE I
CLASSIFICATION PERFORMANCE ON THE HOUSTON2013 DATASET.

Class	Single-Branch					Dual-Branch		
	SpectralFormer	GAHT	GSCViT	morphFormer	DSFormer	DBDA	DBCTNet	Ours
Healthy grass	89.06 ± 6.03	95.67 ± 1.50	94.92 ± 2.32	96.92 ± 1.44	95.32 ± 1.85	93.27 ± 3.28	93.89 ± 4.27	96.37 ± 2.19
Stressed grass	94.09 ± 6.08	99.40 ± 0.37	99.17 ± 0.55	98.65 ± 2.16	99.53 ± 0.26	96.61 ± 3.38	96.43 ± 4.48	99.15 ± 1.05
Synthetic grass	99.07 ± 0.65	99.29 ± 0.33	99.32 ± 0.36	99.17 ± 0.68	99.23 ± 0.24	98.79 ± 0.38	99.04 ± 0.49	99.17 ± 0.30
Trees	95.11 ± 2.39	99.20 ± 0.76	98.94 ± 0.32	97.76 ± 1.20	98.41 ± 2.19	98.86 ± 0.36	98.64 ± 0.61	99.40 ± 0.51
Soil	99.97 ± 0.07	99.68 ± 0.64	99.98 ± 0.03	99.56 ± 0.30	99.85 ± 0.30	100.00 ± 0.00	99.85 ± 0.30	99.70 ± 0.40
Water	97.09 ± 1.20	99.20 ± 0.99	99.56 ± 0.87	100.00 ± 0.00	99.93 ± 0.15	99.56 ± 0.71	99.56 ± 0.87	99.56 ± 0.87
Residential	87.83 ± 3.08	96.70 ± 1.64	94.55 ± 2.37	95.39 ± 2.19	95.73 ± 3.93	95.94 ± 1.61	94.47 ± 3.20	97.45 ± 1.31
Commercial	71.11 ± 5.81	93.07 ± 2.74	90.70 ± 3.22	90.20 ± 4.09	92.29 ± 3.11	90.69 ± 2.53	90.28 ± 4.47	92.65 ± 2.97
Road	84.01 ± 2.20	93.83 ± 1.06	93.89 ± 2.11	91.91 ± 1.94	95.11 ± 0.60	91.25 ± 1.98	91.65 ± 2.60	94.56 ± 1.08
Highway	73.98 ± 8.53	98.18 ± 1.76	98.15 ± 1.25	97.84 ± 1.09	98.79 ± 1.05	99.64 ± 0.16	97.13 ± 1.14	99.17 ± 0.78
Railway	77.03 ± 2.92	96.57 ± 0.79	96.37 ± 1.90	94.11 ± 1.09	94.80 ± 5.36	93.79 ± 4.94	94.28 ± 3.00	96.81 ± 0.50
Parking Lot 1	78.14 ± 4.29	95.40 ± 2.96	95.25 ± 5.67	94.57 ± 4.82	95.38 ± 3.44	93.78 ± 5.84	93.68 ± 6.10	94.76 ± 4.19
Parking Lot 2	88.54 ± 3.09	99.95 ± 0.10	97.66 ± 2.58	99.67 ± 0.44	97.61 ± 1.10	95.51 ± 0.28	95.70 ± 0.15	98.52 ± 1.83
Tennis Court	99.58 ± 0.32	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.95 ± 0.11	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Running Track	100.00 ± 0.00	99.97 ± 0.07	100.00 ± 0.00	99.74 ± 0.34	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
OA	87.06 ± 0.95	97.24 ± 0.34	96.70 ± 0.55	96.33 ± 0.48	96.98 ± 0.73	95.94 ± 0.42	95.66 ± 0.67	97.40 ± 0.32
AA	88.97 ± 0.64	97.74 ± 0.31	97.23 ± 0.47	97.03 ± 0.39	97.46 ± 0.59	96.51 ± 0.36	96.31 ± 0.59	97.82 ± 0.34
Kappa	86.01 ± 1.03	97.02 ± 0.36	96.43 ± 0.59	96.03 ± 0.52	96.73 ± 0.79	95.60 ± 0.45	95.30 ± 0.72	97.19 ± 0.48
train time	27.54 ± 0.45	83.75 ± 10.27	32.54 ± 1.08	152.70 ± 1.40	177.53 ± 2.84	62.89 ± 0.43	137.18 ± 1.66	79.80 ± 1.80
test time	11.28 ± 0.32	81.08 ± 1.31	31.12 ± 0.36	190.12 ± 1.00	302.70 ± 0.99	57.95 ± 1.40	54.04 ± 2.11	106.14 ± 0.87

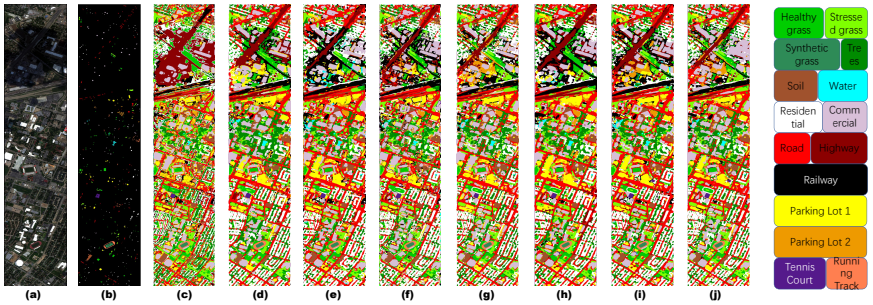


Fig. 2. Visual comparison of different models on the Houston2013 dataset. (a) False-color image. (b) Ground truth. (c) SpectralFormer. (d) GAHT. (e) GSCViT. (f) MorphFormer. (g) DSFormer. (h) DBDA. (i) DBCTNet. (j) DSFMNet.

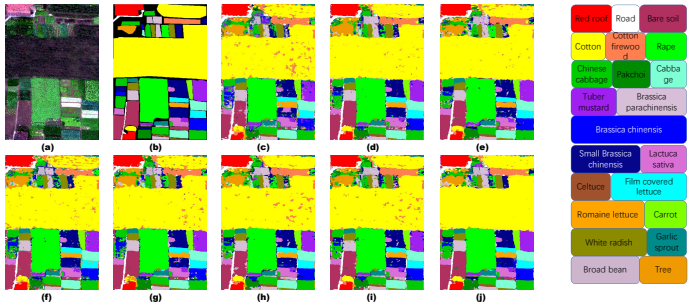


Fig. 3. Visual comparison of different models on the Whu-Hi-HongHu dataset. (a) False-color image. (b) Ground truth. (c) SpectralFormer. (d) GAHT. (e) GSCViT. (f) MorphFormer. (g) DSFormer. (h) DBDA. (i) DBCTNet. (j) DSFMNet.

III. EXPERIMENT

A. Datasets

We conducted experiments on the following three datasets:

1) Houston2013: Acquired by the ITRES CASI-1500 sensor over the University of Houston and its surrounding areas, this dataset contains 144 spectral bands ranging from 380 nm to 1050 nm, with a spatial resolution of 2.5 m. The scene dimensions are 349×1905 pixels, comprising 15 land cover classes.

2) Whu-Hi-HongHu: Collected in Honghu City, China, by Wuhan University using a Headwall Nano-Hyperspec sensor, this dataset comprises 270 spectral bands (400–1000 nm) with a high spatial resolution of 0.109 m. The 940×475 pixel scene contains 22 fine-grained land cover categories.

3) Houston2018: This dataset integrates co-registered hy-

TABLE II
CLASSIFICATION PERFORMANCE ON THE WHU-HI-HONGHU DATASET.

Class	Single-Branch					Dual-Branch		
	SpectralFormer	GAHT	GSCViT	morphFormer	DSFormer	DBDA	DBCTNet	Ours
Red roof	89.68 ± 4.15	95.42 ± 0.89	95.75 ± 0.53	94.16 ± 1.62	96.08 ± 0.62	94.41 ± 0.64	94.13 ± 1.78	96.00 ± 0.92
Road	93.64 ± 1.80	95.17 ± 1.11	96.04 ± 1.15	93.78 ± 2.83	96.37 ± 1.68	96.03 ± 1.03	94.21 ± 1.56	96.33 ± 1.73
Bare soil	80.01 ± 2.36	89.75 ± 1.39	88.80 ± 5.27	90.56 ± 1.71	93.33 ± 0.67	83.82 ± 1.59	88.60 ± 2.62	94.60 ± 0.88
Cotton	95.06 ± 1.20	97.19 ± 1.40	97.72 ± 0.95	95.76 ± 1.14	96.87 ± 0.98	96.23 ± 1.37	96.79 ± 1.16	97.51 ± 0.98
Cotton firewood	95.75 ± 1.21	93.55 ± 4.44	92.84 ± 1.48	94.21 ± 1.71	96.31 ± 1.26	95.27 ± 1.39	95.46 ± 1.38	97.43 ± 0.84
Rape	91.67 ± 1.78	94.30 ± 0.78	96.79 ± 1.32	93.58 ± 2.16	96.18 ± 0.51	93.33 ± 2.12	93.37 ± 1.54	97.06 ± 0.88
Chinese cabbage	72.48 ± 3.46	81.59 ± 7.04	82.39 ± 2.51	82.54 ± 3.00	84.85 ± 2.34	80.37 ± 4.39	82.42 ± 2.12	86.61 ± 2.42
Pakchoi	74.75 ± 4.47	85.12 ± 2.20	91.59 ± 1.69	86.96 ± 3.75	92.90 ± 2.17	81.41 ± 1.86	86.69 ± 2.39	93.41 ± 1.40
Cabbage	94.74 ± 0.93	96.63 ± 1.57	96.90 ± 1.60	96.01 ± 1.62	97.21 ± 1.77	95.93 ± 1.61	95.80 ± 1.12	97.08 ± 1.42
Tuber mustard	74.26 ± 3.82	91.13 ± 2.23	93.33 ± 1.84	89.08 ± 2.48	92.95 ± 1.16	89.65 ± 1.28	89.09 ± 1.53	92.55 ± 1.10
Brassica parachinensis	77.04 ± 2.56	86.29 ± 10.45	91.38 ± 1.53	88.19 ± 3.03	93.41 ± 1.89	87.96 ± 4.75	89.36 ± 1.79	95.18 ± 1.47
Brassica chinensis	66.95 ± 3.45	84.11 ± 3.88	87.73 ± 2.71	83.04 ± 3.09	87.52 ± 3.33	77.50 ± 3.40	80.58 ± 2.65	87.54 ± 3.17
Small Brassica chinensis	68.15 ± 2.37	83.42 ± 1.91	79.54 ± 2.71	77.95 ± 4.58	83.28 ± 1.83	75.66 ± 2.21	77.76 ± 3.30	83.92 ± 0.79
Lactuca sativa	90.16 ± 2.79	94.30 ± 2.01	96.51 ± 1.20	94.54 ± 2.44	97.66 ± 0.81	93.64 ± 1.05	94.00 ± 1.62	98.45 ± 0.56
Celtuce	97.88 ± 0.89	99.33 ± 0.94	99.56 ± 0.28	99.29 ± 0.78	99.35 ± 0.50	99.39 ± 0.35	98.66 ± 0.85	99.12 ± 0.83
Film covered lettuce	90.50 ± 1.40	97.97 ± 0.99	97.05 ± 2.19	94.68 ± 1.84	96.49 ± 1.77	97.82 ± 0.66	96.32 ± 1.28	97.55 ± 1.61
Romaine lettuce	86.22 ± 7.26	98.30 ± 0.84	98.57 ± 1.52	95.95 ± 1.42	97.76 ± 1.81	97.20 ± 1.77	96.72 ± 2.32	98.30 ± 1.20
Carrot	97.36 ± 0.54	96.99 ± 1.93	97.12 ± 1.65	98.07 ± 0.55	99.41 ± 0.27	98.61 ± 0.47	98.64 ± 0.52	99.41 ± 0.32
White radish	94.51 ± 1.05	94.18 ± 1.70	94.56 ± 1.48	93.26 ± 1.51	96.44 ± 0.78	95.26 ± 0.83	95.77 ± 0.64	97.09 ± 1.10
Garlic sprout	94.74 ± 0.51	96.33 ± 1.56	97.69 ± 0.82	96.56 ± 1.57	98.59 ± 0.63	97.14 ± 0.52	97.94 ± 0.47	98.95 ± 0.24
Broad bean	94.15 ± 5.63	98.75 ± 0.75	97.17 ± 4.20	98.39 ± 1.51	99.36 ± 0.34	99.81 ± 0.24	99.08 ± 0.83	99.77 ± 0.28
Tree	96.01 ± 1.80	98.82 ± 1.57	97.57 ± 2.24	97.49 ± 2.24	99.05 ± 0.65	97.28 ± 1.97	97.02 ± 2.15	98.97 ± 0.79
OA	88.37 ± 0.45	93.49 ± 1.37	94.19 ± 0.42	92.41 ± 0.39	94.59 ± 0.50	91.91 ± 0.54	92.78 ± 0.29	95.32 ± 0.38
AA	87.08 ± 0.35	93.12 ± 1.23	93.94 ± 0.38	92.46 ± 0.25	95.01 ± 0.21	91.99 ± 0.49	92.66 ± 0.43	95.58 ± 0.22
Kappa	85.47 ± 0.52	91.83 ± 1.69	92.69 ± 0.52	90.49 ± 0.47	93.24 ± 0.59	89.86 ± 0.66	90.94 ± 0.36	94.11 ± 0.47
train time	51.80 ± 2.87	119.28 ± 1.79	77.43 ± 1.63	226.00 ± 0.47	264.54 ± 2.53	98.10 ± 2.39	200.35 ± 1.14	129.77 ± 1.49
test time	9.23 ± 0.24	54.32 ± 0.19	38.10 ± 0.17	127.75 ± 0.22	207.12 ± 2.25	39.68 ± 0.25	35.91 ± 0.63	70.81 ± 0.17

Fig. 4. Visual comparison of different models on the Houston2018 dataset. (a) False-color image. (b) Ground truth. (c) SpectralFormer. (d) GAHT. (e) GSCViT. (f) MorphFormer. (g) DSFormer. (h) DBDA. (i) DBCTNet. (j) DSFMNet.

perspectral imagery (48 bands, 380–1050 nm). It covers the University of Houston campus and neighboring urban environments with a spatial resolution of 1 m. The scene spans 601×2384 pixels and features 20 land cover classes.

B. Experimental Setup

All experiments are conducted on a platform equipped with an NVIDIA GeForce RTX 4060 Laptop GPU(16GB).

Following DSFormer [29], we select 50 training samples per class for Houston2013 and Whu-Hi-HongHu, and 100 for Houston2018 (reduced to 15 for classes with insufficient samples). All models are trained for 300 epochs with a batch size of 256 using the AdamW optimizer (learning rate: 1×10^{-4}, weight decay: 1×10^{-5}). Inputs are PCA-reduced to 30 spectral bands and extracted as 10×10 patches. Baselines employ their official implementations and default hyperparameters. Training time encompasses the complete 300-epoch pipeline (including PCA), whereas testing time reflects a single sliding-window inference pass (excluding pre-computed PCA). All reported results and time metrics are averaged across five independent runs. Performance is evaluated using Overall Accuracy (OA), Average Accuracy (AA), and the Kappa coefficient. For all three metrics, the values reported in %, where higher is better.

C. Quantitative Evaluation

We compare DSFMNet with seven state-of-the-art HSI classification methods on the Houston2013, Whu-Hi-HongHu, and Houston2018 datasets. These baselines include SpectralFormer [18], GAHT [19], GSCViT [30], MorphFormer [31], two typical dual-branch methods: DBDA [32] and DBCTNet [33], and a recent representative model (2025), DSFormer [29]. The results are presented in Tables I to III. The quantitative results include the prediction accuracy for each class, with the overall performance evaluated by OA, AA, and Kappa. The best results are highlighted in red, and the second-best results are marked in blue.

TABLE III
CLASSIFICATION PERFORMANCE ON THE HOUSTON2018 DATASET.

Class	Single-Branch					Dual-Branch		
	SpectralFormer	GAHT	GSCViT	morphFormer	DSFormer	DBDA	DBCTNet	Ours
Healthy grass	92.48 ± 2.81	88.99 ± 4.94	93.23 ± 3.38	90.82 ± 2.56	89.43 ± 7.08	92.41 ± 1.54	91.38 ± 2.39	90.23 ± 7.84
Stressed grass	81.46 ± 1.79	81.81 ± 5.55	82.98 ± 4.39	85.05 ± 1.63	87.42 ± 5.80	85.40 ± 2.19	84.88 ± 4.28	84.98 ± 3.12
Artificial turf	99.55 ± 0.17	100.00 ± 0.00	99.83 ± 0.11	99.76 ± 0.17	99.73 ± 0.17	100.00 ± 0.00	99.97 ± 0.07	99.93 ± 0.08
Evergreen trees	96.62 ± 0.59	96.41 ± 1.16	96.33 ± 0.84	95.65 ± 1.52	95.50 ± 2.06	97.19 ± 0.95	96.54 ± 1.05	97.47 ± 0.94
Deciduous trees	89.09 ± 1.72	95.07 ± 0.94	93.13 ± 1.71	93.35 ± 0.89	93.92 ± 2.15	93.59 ± 1.94	91.73 ± 2.10	94.70 ± 1.07
Bare earth	98.38 ± 2.23	99.89 ± 0.14	99.55 ± 0.27	99.47 ± 0.67	99.69 ± 0.28	99.89 ± 0.23	99.95 ± 0.06	99.80 ± 0.19
Water	99.76 ± 0.48	100.00 ± 0.00	99.76 ± 0.48	100.00 ± 0.00	100.00 ± 0.00	99.88 ± 0.24	100.00 ± 0.00	100.00 ± 0.00
Residential buildings	80.54 ± 1.26	85.76 ± 1.33	83.91 ± 3.24	82.63 ± 0.93	89.23 ± 2.14	89.83 ± 1.11	81.10 ± 8.81	88.89 ± 2.75
Non-residential buildings	67.90 ± 3.57	79.09 ± 2.37	80.23 ± 1.91	78.14 ± 1.98	83.71 ± 2.09	82.84 ± 1.30	84.47 ± 5.87	86.39 ± 1.27
Roads	28.03 ± 8.31	52.34 ± 3.83	47.61 ± 3.90	47.14 ± 4.18	50.80 ± 3.96	53.61 ± 3.14	45.24 ± 9.68	53.92 ± 3.48
Sidewalks	33.69 ± 5.60	57.56 ± 4.48	53.78 ± 3.63	47.66 ± 1.68	61.91 ± 6.94	57.45 ± 3.65	64.41 ± 7.30	56.98 ± 2.70
Crosswalks	53.31 ± 6.46	81.29 ± 3.60	79.03 ± 4.59	73.79 ± 3.31	83.46 ± 5.07	79.03 ± 3.90	73.09 ± 8.17	82.18 ± 1.56
Major thoroughfares	52.59 ± 8.27	70.19 ± 6.23	67.28 ± 6.98	64.37 ± 4.76	69.53 ± 5.15	69.19 ± 2.61	60.77 ± 8.85	73.66 ± 6.20
Highways	93.14 ± 3.76	96.80 ± 1.07	97.11 ± 1.33	95.92 ± 1.26	97.48 ± 1.44	97.78 ± 1.56	95.49 ± 2.66	97.02 ± 1.24
Railways	98.68 ± 0.97	98.58 ± 0.86	98.51 ± 0.61	97.55 ± 1.85	98.56 ± 0.59	98.10 ± 0.76	97.38 ± 0.53	98.86 ± 0.79
Paved parking lots	88.45 ± 1.69	91.85 ± 2.97	93.50 ± 1.49	89.15 ± 1.57	91.40 ± 3.80	91.43 ± 1.46	87.57 ± 6.96	95.02 ± 2.02
Unpaved parking lots	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Cars	86.47 ± 3.25	92.96 ± 2.95	93.26 ± 2.00	93.89 ± 0.53	96.44 ± 1.42	93.54 ± 0.98	92.68 ± 5.20	94.15 ± 1.22
Trains	94.70 ± 1.28	98.59 ± 1.03	97.69 ± 2.31	97.41 ± 0.78	99.22 ± 0.48	99.02 ± 0.81	98.77 ± 1.02	99.16 ± 0.78
Stadium seats	98.49 ± 1.02	99.68 ± 0.32	99.45 ± 0.48	98.64 ± 0.95	99.07 ± 1.20	99.62 ± 0.58	99.37 ± 0.36	99.64 ± 0.43
OA	66.47 ± 1.60	77.68 ± 0.90	77.25 ± 1.28	75.41 ± 0.50	80.49 ± 0.91	80.00 ± 0.54	78.70 ± 0.83	81.95 ± 0.64
AA	81.67 ± 0.52	88.34 ± 0.30	87.81 ± 0.48	86.52 ± 0.15	89.32 ± 0.62	88.99 ± 0.20	87.24 ± 1.17	89.65 ± 0.38
Kappa	59.55 ± 1.60	72.29 ± 1.00	71.67 ± 1.46	69.51 ± 0.50	75.52 ± 1.03	74.95 ± 0.60	73.19 ± 0.69	77.19 ± 0.77
train time	74.15 ± 2.71	208.37 ± 0.53	99.14 ± 2.26	421.13 ± 1.25	486.35 ± 0.33	179.31 ± 2.48	360.12 ± 6.99	212.09 ± 1.40
test time	25.29 ± 0.73	169.57 ± 0.77	72.47 ± 2.03	412.78 ± 1.35	652.65 ± 1.16	127.98 ± 1.05	118.72 ± 9.55	223.50 ± 1.09

1) Accuracy Analysis: DSFMNet achieves superior performance across all evaluation indicators. On Houston2013, our method effectively handles complex urban structures and shadow occlusion, yielding the highest OA (97.40%) and demonstrating robustness to class imbalance with a leading AA (97.82%). The performance gains on Whu-Hi-HongHu (95.32% OA) are primarily attributed to CF-FMViM's ability to extract fine-grained representations. Notably, on the challenging Houston2018 dataset, DSFMNet outperforms DSFormer by 1.46% in OA, validating the synergistic integration of FM-TSSA's global semantics and CF-FMViM's local feature refinement.

Compared with prior dual-branch networks (e.g., DBDA and DBCTNet) that do not exploit frequency-domain cues and thus exhibit redundant representations (Houston2018 OAs of 80.00% and 78.70%), our design incorporates frequency modulation to promote complementary spatial–frequency interactions, reduce redundancy, and improve discriminability, leading to consistent accuracy gains.

2) Efficiency Analysis: Beyond accuracy, DSFMNet exhibits exceptional computational efficiency. By leveraging linear-complexity $\mathcal{O}(n)$ paradigms—specifically Vision Mamba and linear attention—coupled with strategic downsampling, DSFMNet achieves a $\sim$41% speedup over the dual-branch DBCTNet (212.09s vs. 360.12s on Houston2018). Notably, despite its dual-branch architecture, DSFMNet operates over $2\times$ faster than single-branch Transformers such as DSFormer (486.35s) and MorphFormer (421.13s). These results confirm that a meticulously designed dual-branch hybrid architecture can effectively circumvent the heavy computational overhead typical of standard self-attention.

D. Ablation Study

We conduct ablations on all three datasets to evaluate the dual-branch design and key modules in DSFMNet. Table V removes the fine/coarse branch, while Table IV replaces DSFormer-style components with FM-TSSA/CF-FMViM.

1) Effectiveness of the Dual-Branch Architecture: Individual branch evaluations reveal that TSSA-FM underperforms on Whu-Hi-HongHu (93.21% OA) due to insufficient discriminability for spectrally similar crops, while CF-FMViM improves results to 95.10% by capturing fine-grained textures. On Houston2018, single-branch performance plateaus at 80.97% OA. Their integration into the full DSFMNet boosts OA to 81.95%, demonstrating that the synergy between coarse-scale global semantics and fine-scale frequency details is essential for superior performance.

2) Impact of Key Modules on Accuracy and Efficiency: We evaluate accuracy-speed trade-offs by progressively replacing Baseline [29] components (Table IV). Replacing self-attention with FM-TSSA reduces Houston2018 training time by over $3\times$ (499.44s to 158.08s), confirming its $\mathcal{O}(n)$ efficiency. Incorporating CF-FMViM restores OAs to 81.95% (Houston2018) and 95.32% (Whu-Hi-HongHu), proving it compensates for linear attention's information loss via high-resolution frequency modeling, thereby achieving optimal accuracy with minimal computational overhead.

IV. CONCLUSION

This paper introduces DSFMNet, a dual-scale frequency-modulated network with linear-complexity hybrid architecture for hyperspectral image classification. We demonstrate that architectural efficiency and accuracy can coexist through carefully designed asymmetric pathways with complementary

TABLE IV
COMPONENT REPLACEMENT ABLATION UNDER A FIXED DUAL-BRANCH BACKBONE (✓: REPLACING THE CORRESPONDING MODULE)

Modules		Houston2013			Whu-Hi-HongHu			Houston2018		
FM-TSSA	CF-FMViM	OA (%)	Train (s)	Test (s)	OA (%)	Train (s)	Test (s)	OA (%)	Train (s)	Test (s)
×	×	97.32 ± 0.38	185.84 ± 2.95	320.23 ± 3.92	94.21 ± 0.66	279.71 ± 6.14	220.22 ± 8.88	80.52 ± 0.62	499.44 ± 7.99	702.82 ± 21.47
✓	×	97.50 ± 0.45	63.69 ± 3.78	76.42 ± 1.16	95.07 ± 0.30	92.13 ±1.52	40.68 ± 0.07	80.32 ± 0.99	158.08 ± 9.52	130.62 ± 0.39
✓	✓	97.40 ± 0.32	79.80 ± 1.80	106.14 ± 0.87	95.32 ±0.38	129.77 ±1.49	70.81 ± 0.17	81.95 ±0.64	212.13 ± 2.21	228.93 ±3.20

TABLE V
STRUCTURE ABLATION ON THREE DATASETS.

Branches		Overall Accuracy (%)		
Coarse branch	Fine branch	Houston2013	Whu-Hi-HongHu	Houston2018
✓	×	97.17 ± 0.61	93.21 ± 2.66	80.15 ± 0.93
×	✓	97.04 ± 0.72	95.10 ± 0.67	80.97 ± 0.83
✓	✓	97.40 ± 0.32	95.32 ± 0.38	81.95 ± 0.64

functions. This approach extends to other multi-scale vision tasks, opening new research directions for efficient multi-scale processing.

V. ACKNOWLEDGMENT

This work was supported by the Research Foundation of Education Bureau of Hunan Province under Grant 25C0086.

REFERENCES

[1] V. Kumar, R. S. Singh, M. Rambabu, and Y. Dua, "Deep learning for hyperspectral image classification: A survey," *Comput. Sci. Rev.*, vol. 53, Art. no. 100658, 2024.

[2] L. Bai, Z. Chen, and W. Zhang, "HSI classification: A comprehensive review, dataset, and future trends," *J. Indian Soc. Remote Sens.*, vol. 53, no. 12, pp. 3983–4004, 2025.

[3] N. Anand and A. Arulmurugan, "Current advances in hyperspectral remote sensing in urban planning," in *Proc. 3rd Int. Conf. Intell. Comput. Instrum. Control Technol. (ICICICT)*, Aug. 2022, pp. 94–98.

[4] Y. Song *et al.*, "Deep learning for hyperspectral image classification: A comprehensive review and future predictions," *Inf. Fusion*, vol. 123, Art. no. 103285, 2025.

[5] S. Mastrantonis *et al.*, "A novel method for robust marine habitat mapping using a kernelised aquatic vegetation index," *ISPRS J. Photogramm. Remote Sens.*, vol. 209, pp. 472–480, 2024.

[6] S. Ren et al., "Desorption kinetics of tetracyclines in soils assessed by diffusive gradients in thin films," Environ. Pollut., vol. 256, Art. no. 113394, Jan. 2020.

[7] S. Lu *et al.*, "SSUM: Spatial–spectral unified Mamba for hyperspectral image classification," *Remote Sens.*, vol. 16, no. 24, Art. no. 4653, Dec. 2024.

[8] S. Lai et al., "Predicting lysine phosphoglycerylation sites using bidirectional encoder representations with transformers and protein feature extraction and selection," in 2022 15th Int. Congr. Image Signal Process. BioMed. Eng. Inf. (CISP-BMEI), 2022, pp. 1–6.

[9] Y. Zhong, X. Wang, S. Wang, and L. Zhang, "Advances in spaceborne hyperspectral remote sensing in China," Geo-Spat. Inf. Sci., vol. 24, no. 1, pp. 95–120, Mar. 2021.

[10] R. Wang, X. Ye, Y. Huang, M. Ju, and W. Xiang, "GASSF-Net: Geometric algebra based spectral-spatial hierarchical fusion network for hyperspectral and LiDAR image classification," Remote Sens., vol. 16, no. 20, Art. no. 3825, Oct. 2024.

[11] 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, 2023.

[12] J. Li, B. Xi, Y. Li, Q. Du, and K. Wang, "Hyperspectral classification based on texture feature enhancement and deep belief networks," Remote Sens., vol. 10, no. 3, Art. no. 396, Mar. 2018.

[13] R. Ramo and E. Chuvieco, "Developing a random forest algorithm for MODIS global burned area classification," Remote Sens., vol. 9, no. 11, Art. no. 1193, Nov. 2017.

[14] Y. Zhong, T. Jia, J. Zhao, X. Wang, and S. Jin, "Spatial-spectral-emissivity land-cover classification fusing visible and thermal infrared hyperspectral imagery," Remote Sens., vol. 9, no. 9, Art. no. 910, Sep. 2017.

[15] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," J. Sensors, vol. 2015, Art. no. 258619, 2015.

[16] D. Hong et al., "Graph convolutional networks for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

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

[18] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–15, 2022.

[19] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–14, 2022.

[20] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv preprint arXiv:2312.00752, Dec. 2023.

[21] W. Zhou, S. Kamata, H. Wang, M. S. Wong, and H. Hou, "Mamba-in-Mamba: Centralized Mamba-cross-scan in tokenized Mamba model for hyperspectral image classification," Neurocomputing, vol. 613, Art. no. 128751, Nov. 2025.

[22] R. J. Marks II, Handbook of Fourier Analysis and Its Applications. Oxford, U.K.: Oxford Univ. Press, 2009.

[23] C. Debes et al., "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[24] Y. Zhong et al., "WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," Remote Sens. Environ., vol. 250, Art. no. 112012, Dec. 2020.

[25] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, 2019.

[26] S. Lee, J. Choi, and H. J. Kim, "EfficientViM: Efficient Vision Mamba with hidden state mixer based state space duality," arXiv preprint arXiv:2411.15241, Nov. 2024.

[27] H. Li et al., "Frequency-aware transformer for learned image compression," arXiv preprint arXiv:2310.16387, Oct. 2023.

[28] Z. Wu *et al.*, "Token statistics transformer: Linear-time attention via variational rate reduction," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[29] Y. Xu, D. Wang, L. Zhang, and L. Zhang, "Dual selective fusion transformer network for hyperspectral image classification," Neural Netw., vol. 187, Art. no. 107311, 2025.

[30] Z. Zhao and X. Xu, "A lightweight network of groupwise separable convolution and vision transformer for hyperspectral image classification," in *Proc. SPIE*, vol. 13179, paper 1317917, 2024.

[31] S. K. Roy et al., "Spectral–spatial morphological attention transformer for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–15, 2023.

[32] R. Li, S. Zhang, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," Remote Sens., vol. 12, no. 3, Art. no. 582, Feb. 2020.

[33] R. Xu et al., "DBCTNet: Double branch convolution-transformer network for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens., vol. 62, pp. 1–15, 2024.