DTM: DCT–Transformer–Mamba Architecture for Improving Hyperspectral Image Classification

Weijia Cao, *Member, IEEE*, Xiaofei Yang, *Member, IEEE*, Yicong Zhou, *Senior Member, IEEE*, Fu Wang, *Member, IEEE*, and Xiang Zhou

Abstract*—Hyperspectral image (HSI) classification is faced with significant challenges due to the nonuniform spatial self-similarity, spectral instability, and high dimensionality of the data. Existing methods typically focus on adjusting model architectures to uncover deep dependencies within HSI. However, these approaches often struggle with the inconsistencies in spatial dependencies, particularly in the context of long-range dependencies across heterogeneous regions. In this article, we propose a novel discrete cosine transform (DCT)–Transformer–Mamba (DTM) architecture, which introduces a new perspective by transforming irregular, unbalanced spatial data into a more ordered and consistent form to enhance long-range dependencies. Our main innovation lies in using autocorrelation function (ACF) analysis to identify spatial dependence patterns that hinder effective modeling. To address this, we introduce a DCT-augmented processing module (DCTConv), powered by harmonic neural networks (HNNs), which extends spatial correlation range by 13.57% and improves spatial consistency, thereby stabilizing long-range dependencies. This process enhances the ability of Mamba to capture complex spatial dependencies across different regions. In addition, to mitigate the inherent 1-D sequence modeling bias of Mamba, we incorporate DCT-Scanning Mamba blocks with hierarchical skip connections. The DTM architecture thus better captures directional-sensitive features, leading to a 56.7% improvement in performance for these classes on the Indian Pines dataset. The DTM framework integrates DCT-based frequency-domain regularization, Mamba, and Transformer modules, resulting in a 99% reduction in parameters compared to HSI Transformer baselines and 50% faster training times compared to Mamba-based models. Extensive experiments on benchmark datasets demonstrate the effectiveness of DTM, achieving a state of the art 98.76% overall accuracy (OA) on the Kennedy Space Center dataset, with a 6.69% improvement in low-sample regimes over Mamba baselines. The code for this work will be made publicly available at https://github.com/xiachangxue/DeepHyperX

***Index Terms*—Discrete cosine transform, hyperspectral image (HSI) classification, Mamba, Transformer.**

Received 19 August 2025; revised 28 December 2025; accepted 29 January 2026. Date of publication 6 February 2026; date of current version 20 February 2026. This work was supported by the Key Research and Development Program of Shandong Province, China under Project 2025CXGC010113. *(Corresponding authors: Xiaofei Yang; Xiang Zhou.)*

Weijia Cao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: caowj@aircas.ac.cn).

Xiaofei Yang is with Guangzhou University, Guangzhou 510006, China (e-mail: xiaofeiyang@gzhu.edu.cn).

Yicong Zhou is with the University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo).

Fu Wang is with the CMA Earth System Modeling and Prediction Centre, Beijing 100081, China, and also with the State Key Laboratory of Severe Weather Meteorological Science and Technology, Beijing 100081, China (e-mail: wangfu@cma.cn).

Xiang Zhou is with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: zhouxiang@aircas.ac.cn).

Digital Object Identifier 10.1109/TGRS.2026.3661532

I. Introduction

HYPERSPECTRAL image (HSI) is a vital remote sensing technique, providing detailed spectral information across numerous bands for applications in environmental monitoring, agriculture, and military surveillance [1]. However, HSI data pose significant challenges for accurate classification due to their high dimensionality, inherent redundancy, and noise, which complicate the extraction of meaningful features [2], [3], [4]. In particular, HSI data exhibit.

1) *Nonuniform Spatial Self-Similarity:* While HSI exhibits strong spatial autocorrelation patterns, these are often disrupted by sensor artifacts and illumination variations, creating heterogeneous decay rates in spatial dependencies (see Fig. 1) [5], [6]. As explicitly shown in Fig. 1, the raw data variance is significant, whereas our proposed processing significantly narrows this distribution, achieving spatial dependency homogenization that is critical for robust modeling.
2) *Spectral Instability:* Subtle atmospheric effects and noise disproportionately distort spectral signatures while maintaining spatial structures [7].
3) *High-Dimensional Entanglement:* High-dimensional data redundancy, leading to computational inefficiencies and difficulties in effective modeling [8], [9].

Traditional convolutional neural networks (CNNs) have demonstrated effectiveness in capturing local spatial features [10], yet their fixed receptive fields limit their ability to extract global spectral context [11]. Transformer models, employing self-attention mechanisms, excel at modeling long-range dependencies across spectral bands, but they require large-scale labeled datasets and involve substantial computational overhead [12], [13]. While CNNs and Transformers excel in spectral and local spatial feature extraction, they are not well-suited for handling complex spatial dependencies and large-scale long-range interactions in high-dimensional HSI data [11], [14].

In contrast, Mamba models, based on state-space model (SSM), have shown potential in capturing both local and long-range dependencies in sequence data. SSMs mitigate this by using linear differential equations to update hidden states, thus avoiding the nonlinear instability common in recurrent neural networks (RNNs). This ensures stable gradient propagation, making SSMs well-suited for handling long-range dependencies in complex data. In addition, Mamba

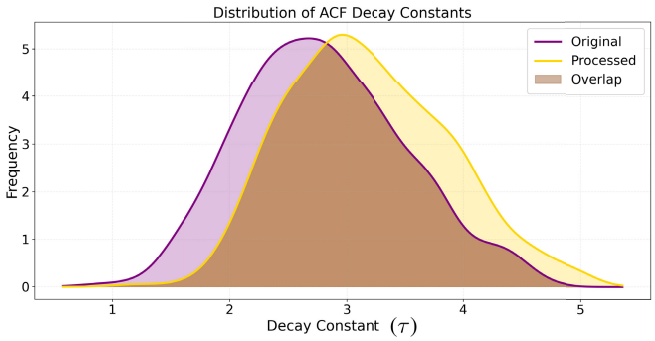

Fig. 1. Distribution of spatial autocorrelation decay constants (τ) on the IPs dataset. The density plots compare the statistical spread of decay constants for the original HSI data (purple) and the processed features (yellow). The original distribution is broad and left-skewed, indicating inconsistent spatial dependencies. The processed distribution shifts rightward and becomes significantly narrower, demonstrating that DCT-augmented processing effectively homogenizes spatial coherence and enhances structural stability. When both of the first two have values in the same range, this part appears as a mixture of the two colors (brown).

and other SSM-based models like S4 incorporate Fourier transforms to smooth gradient flows, further reducing gradient-related issues and enhancing the modeling of long-range dependencies in both temporal and spatial sequences [15]. However, these models encounter issues with unstable or uncertain sequences, leading to problems such as gradient vanishing or explosion, particularly in long sequences [16]. Mamba, originally designed for 1-D data, faces limitations when applied to high-dimensional data like HSI. It struggles when processing 2-D images due to the anisotropic nature of spatial dependencies, and its 1-D scanning bias creates directional preference artifacts when applied to 2-D data [17]. Simply adjusting the scanning direction does not fully capture the complex spatial and spectral interactions inherent in HSI data, which affects Mamba's performance in these scenarios. To address these issues, modifications such as HyperMamba have been proposed, which adaptively scan spatial and spectral information, improving the interaction between spatial and spectral features during processing. These models enhance HSI classification performance while maintaining the computational efficiency of the original Mamba model [18].

Our key insight stems from a systematic analysis of the autocorrelation function (ACF) to investigate the spatial structure of HSIs. We observe that raw HSI data exhibit heterogeneous ACF decay rates across different land cover types (see Fig. 1), which hinder Mamba's ability to model consistent long-range dependencies. To address this, we apply harmonic neural networks (HNNs) (DCTConv) with discrete cosine transform (DCT) preprocessing to achieve ACF decay homogenization, aligning spatial autocorrelation patterns with Mamba's strengths in sequential modeling. It enhances spatial coherence while resulting in more stable spatial dependencies. Based on these insights, we propose the DCT–Transformer–Mamba (DTM) architecture, which integrates three complementary components: 1) a DCT front-end that performs spectral–spatial harmonic filtering to standardize ACF properties; 2) DCT–Mamba blocks that process the enhanced spatial stream through DCT–Mamba Scan mechanisms; and 3) a decorrelation Transformer branch dedicated to spectral feature refinement. This architecture dynamically fuses spectral and spatial information through learnable DCT-attention gates and addresses spectral redundancy, thereby optimizing both spatial and spectral dependency modeling in HSI classification.

The main contributions of this article are summarized as follows.

1) *ACF-Guided Spatial Consistency Enhancement:* We systematically analyze the ACF to reveal inconsistencies in hyperspectral spatial dependencies, which hinder Mamba's ability to model long-range interactions effectively. We propose a DCT-augmented processing module (DCTConv) that extends the spatial correlation range by 13.57%. It improves spatial consistency, facilitating robust sequential modeling by stabilizing long-range dependencies in hyperspectral images.
2) *Multidirectional Spatial Modeling With DCT-Scanning Mamba:* We introduce DCT–Mamba blocks with hierarchical skip connections to overcome Mamba's inherent 1-D sequence bias, leading to a 56.7% improvement in direction-sensitive classes. This multidirectional spatial modeling captures complex spatial dependencies across heterogeneous regions.
3) *Efficient Hybrid Architecture for Balanced Spectral–Spatial Modeling:* The DTM framework integrates DCT-augmented processing (DCTConv), Mamba, and Transformer modules, leading to a 99% reduction in parameters compared to Transformer baselines and 50% faster training than Mamba-based models. DTM achieves state-of-the-art performance with 95.72% accuracy on the Indian Pines (IPs) dataset, even with limited labeled data.

This article is structured as follows. Section II reviews related work. Section III introduces the proposed DTM architecture and its components. Section IV presents experimental results and analysis. Section V concludes this article and discusses future research directions.

II. Related Work

This section reviews existing research on HSI classification, with a focus on the application of HNNs for frequency-domain analysis, the use of Transformer-based and Mamba-based models for capturing both global and local dependencies, and the role of the ACF as a diagnostic tool for evaluating spatial dependencies.

A. HNNs in Image Classification

HNNs have shown promise in various image classification tasks by utilizing DCT basis functions for frequency-domain analysis [19], [20]. While earlier work focused primarily on RGB images, HNN-based techniques have proven to reduce pixel-level correlations and separate low- and high-frequency components. This processing improves robustness under varying illumination conditions and enhances the representation of structural details while suppressing noise. The success of HNN in RGB image analysis provides a strong foundation for

its application to more complex spectral data, such as hyperspectral images, where spectral and spatial features interact more intricately.

B. Transformer-Based and Mamba-Based Models

Transformer models have gained significant attention for their ability to model long-range dependencies through self-attention mechanisms. These models are particularly useful in capturing global spectral correlations in hyperspectral data [21], [22]. However, they are computationally expensive and rely heavily on large labeled datasets, which can be a limitation in many practical applications [12]. On the other hand, Mamba models, based on SSM, handle long-sequence data more efficiently by using linear differential equations to update hidden states, avoiding the nonlinear instability seen in RNNs [23]. Mamba-based models, such as MiM, SpectralMamba, SSMamba, S^2Mamba, MambaHSI, and others, have shown strong performance in handling long-range dependencies and integrating spectral–spatial features [14], [24], [25], [26]. Despite their strengths, Mamba models still face challenges related to spectral variability and computational efficiency, which highlight the need for further optimization.

To this end, this article proposes a novel DTM architecture, which introduces a frequency-guided enhancement mechanism at the front end of the network. In particular, the DCT-augmented processing stream (DCTConv) projects raw spatial information into the frequency domain, extracts dominant orthogonal components, and enhances spatial coherence through learnable parameters. This stabilized representation reduces spatial–spectral redundancy and facilitates long-range dependency modeling in the subsequent Transformer and Mamba modules. Compared with HyperMamba and SpectralMamba, which focus primarily on architectural adaptations for sequence modeling, DTM provides a unique integration of frequency-domain guidance and SSM, achieving both higher accuracy and lower computational cost, especially in HSI scenarios with limited training samples or complex interband variability.

III. Preliminary: Analyzing Spatial Dependencies With the ACF

A. ACF and Spatial Dependency Analysis

The ACF measures the correlation between values of a stationary process at different spatial lags, quantifying how similar data at one location is to data at another, separated by a specific distance. A process exhibits long memory, or long-range dependence, when its autocovariances decay slowly, often following a power law, leading to a divergent sum of autocovariances [27]. It indicates persistent spatial dependencies over long distances, which is closely related to self-similarity. In remote sensing, understanding these long-range dependencies using the ACF helps capture and model spatial consistency in hyperspectral image (HSI) data.

1) 1-D ACF: In the 1-D case, the ACF at lag k, denoted as $R(k)$, is defined as

$$R(k) = \frac{\mathbb{E}\left[(X_i - \mu)\left(X_{i+k} - \mu\right)\right]}{\sigma^2} \tag{1}$$

where μ is the mean, and σ^2 is the variance. The ACF measures how similar adjacent spatial values are, and how this similarity extends over space. In remote sensing, a slowly decaying ACF indicates strong spatial dependencies, which can improve classification accuracy by incorporating information from distant regions.

2) 2-D ACF: The ACF concept can be extended to two dimensions for image data. For an image $F(x, y)$, the 2-D ACF at spatial lag $(\Delta x, \Delta y)$ is defined as

$$R(\Delta x, \Delta y) = \frac{\mathbb{E}\left[(F(x, y) - \mu)\left(F(x + \Delta x, y + \Delta y) - \mu\right)\right]}{\sigma^2} \tag{2}$$

where μ and σ^2 represent the mean and variance of the image. This 2-D ACF is essential for understanding spatial relationships across images and can provide insights into spatial consistency over larger scales. Previous work in remote sensing and texture analysis has shown that 2-D ACF can effectively characterize spatial consistency and improve classification by capturing long-range spatial patterns [28].

3) ACF Decay in Spatial Dimensions: In both 1-D and 2-D ACFs, the decay constant τ is a crucial parameter, indicating the extent of spatial dependencies. In one dimension, the decay can be modeled as

$$R(k) \sim e^{-k/\tau} \tag{3}$$

where larger τ values suggest stronger long-range dependencies in the spatial data. Inconsistent spatial dependencies, quantified by a high variance in the ACF decay constant τ across an image, degrade the performance of sequential models like Mamba. This observation forms the core motivation for our proposed spatial homogenization strategy.

B. Autocorrelation in Remote Sensing: From Feature Extraction to Spatial Homogenization

ACF-based approaches have been widely applied to hyperspectral image classification and remote sensing, where capturing spatial consistency plays a significant role in feature extraction. Faugeras and Pratt [28] demonstrated the utility of 2-D ACF in capturing spatial texture feature extraction. Moreover, Small [29] explored the complexities of spatial and spectral variability by ACF in remote sensing through high spatial resolution spectral mixture analysis of urban reflectance. It emphasized the use of spectral mixture models to quantify spatial heterogeneity across different urban areas, focusing on the physical characteristics of urban reflectance.

However, unlike these traditional approaches that use ACF primarily for texture feature extraction, we employ ACF analysis as a *diagnostic tool* to uncover the structural limitations of raw HSI data. As demonstrated in our analysis (see Fig. 1), raw HSI data exhibits heterogeneous ACF decay rates, which hinder the effectiveness of long-range dependency modeling. This insight drives the design of our DCT-augmented processing stream (DCTConv), which is explicitly designed not for general-purpose frequency filtering, but for *spatial dependency homogenization*. By reducing the variance of τ, we aim to transform irregular 2-D spatial patterns into a more coherent form suitable for sequential modeling.

IV. DTM Architecture

In this section, we present the proposed DTM architecture as shown in Fig. 2, which comprises three core modules: 1) a DCT-augmented processing stream; 2) a hybrid feature extraction module; and 3) a feature enhancement module. This design introduces a novel frequency-domain processing strategy to address spatial dependency inconsistencies in HSIs.

A. DCT-Augmented Processing Stream: Frequency-Guided Spatial Enhancement

The DCT-augmented processing stream (DCTConv) is designed to mitigate spatial dependency inconsistencies by projecting spatial information into the frequency domain. Using DCT basis functions, DCTConv converts the spatial distribution into spectral patterns, extracts the dominant components in the orthogonal domain, and applies learnable parameters to enhance spatial coherence. In this way, spatial redundancy is transformed into a structured and interpretable spatial frequency energy representation, providing a stable and effective path for spatial coherence modeling. Existing models typically adjust network architectures to adaptively capture long-range dependencies within irregular spatial patterns. In contrast, our approach focuses on transforming the input data itself into a more ordered and consistent representation, thus facilitating long-range dependency modeling.

Given an HSI tensor $X \in \mathbb{R}^{H\times W\times C_{\text{in}}}$, where H and W denote spatial dimensions and C_{in} is the number of spectral channels, we employ DCT basis functions to parameterize frequency-domain filtering.

First, the convolution kernel between input channel l and output channel i is constructed via a DCT basis expansion

$$K_{i,l}(x,y) = \sum_{u=0}^{H-1}\sum_{v=0}^{W-1} w_{i,l,u,v}\,\phi_{u,v}(x,y) \tag{4}$$

where $w_{i,l,u,v}$ are learnable coefficients, $\phi_{u,v}(x,y)$ are DCT basis functions, and (x,y) and represent spatial coordinates, and (u,v) are frequency indices along height and width.

The output feature map Y_i is then computed as

$$Y_i = \sum_{l=0}^{C_{\text{in}}-1} \left(K_{i,l} * X_l\right) \tag{5}$$

where * denotes 2-D convolution, and X_l is the lth input feature map.

The DCT basis functions $\phi_{u,v}(x,y)$ are defined as

$$\phi_{u,v}(x,y) = \beta_u^{(H)}\beta_v^{(W)} \cos\left(\frac{\pi(2x+1)u}{2H}\right)\cos\left(\frac{\pi(2y+1)v}{2W}\right) \tag{6}$$

where

$$\beta_k^{(N)} = \begin{cases} \sqrt{\dfrac{1}{N}}, & k=0 \\ \sqrt{\dfrac{2}{N}}, & k=1,2,\ldots,N-1. \end{cases} \tag{7}$$

B. ACF-Based Theoretical Derivation of Long-Range Dependency Improvement

To theoretically demonstrate that the DCT-augmented processing stream enhances long-range spatial dependencies, we analyze its effect on the ACF of a spatial sequence. In this derivation, each variable is defined explicitly.

Assume that the original spatial sequence $X(i)$, where i indexes spatial location, has an ACF that decays exponentially with spatial lag k, which represents the distance (in pixels) between two points in the spatial domain. The raw ACF is modeled as

$$R_{\text{raw}}(k) = R_0 \exp\left(-\frac{|k|}{\tau_{\text{raw}}}\right) \tag{8}$$

where R_0 is the ACF value at zero lag (i.e., at $k=0$, typically $R_0=1$ after normalization), and τ_{raw} is the decay constant of the raw data, indicating the characteristic spatial correlation length; a larger τ_{raw} implies that spatial dependencies persist over a longer distance.

The DCT-augmented processing stream (DCTConv) applies a harmonic convolution to $X(i)$ using a filter $h(i)$ derived from DCT basis functions. This operation is equivalent to a linear filtering that smooths the data. The output signal $Z(i)$ is given by

$$Z(i) = (h * X)(i) = \sum_j h(j)\, X(i-j) \tag{9}$$

where $h(j)$ is the impulse response of the filter at position j, $h(j)$ is derived from the DCT basis functions and represents the weight applied at offset j, and it is normalized so that $\sum_j h(j) = 1$.

Let $R_h(k)$ be the ACF of the filter $h(i)$, which measures the similarity of the filter to a shifted version of itself. We assume that $R_h(k)$ decays exponentially

$$R_{\text{h}}(k) \approx \exp\left(-\frac{|k|}{\tau_{\text{h}}}\right) \tag{10}$$

where τ_h is the decay constant of the filter, reflecting the effective spatial smoothing length introduced by the DCT-based processing, and $|k|$ denotes the absolute lag.

The ACF of the processed signal $Z(i)$ is

$$R_Z(k) = \sum_s R_h(s)\, R_{\text{raw}}(k-s) \tag{11}$$

where the summation variable s runs over the support of the filter. This equation implies that $R_Y(k)$ is a processed version of the raw ACF, weighted by the self-similarity of the filter.

When two exponential decays are convolved, the result is approximately an exponential decay with an effective decay constant given by

$$\tau_{\text{proc}} \approx \tau_{\text{raw}} + \tau_h \tag{12}$$

where τ_{proc} is the effective decay constant of the processed signal $Z(i)$, indicating that the harmonic convolution implemented via DCT basis functions effectively increases the spatial correlation length. In other words, spatial correlations persist over larger distances after processing, thereby enhancing long-range dependencies.

In summary, by performing harmonic convolution with DCT basis functions, the DCT-augmented processing stream

increases the effective ACF decay constant from τ_{raw} to $\tau_{\text{proc}} \approx \tau_{\text{raw}} + \tau_h$. This extension of the spatial correlation length provides a more stable foundation for subsequent modules in the DTM architecture, ultimately contributing to improved performance in hyperspectral image classification while providing a theoretically interpretable mechanism for spatial dependency enhancement, which supports trustworthy model behavior.

C. Hybrid Feature Extraction Module: Spatial–Spectral Interaction

This module combines two complementary models: Mamba for spatial dependency extraction and Transformer for global spectral analysis, as illustrated in Fig. 3.

1) Mamba Branches: The Mamba component employs structured SSM to capture spatial dependencies. The formulation is defined as follows:

$$X_{\text{norm}} = \text{LayerNorm}\,(X_{\text{in}}) \tag{13}$$

$$X_{\text{scan}} = \text{SiLU}(\text{DCT-Mamba}\,(\text{DCTConv}\,(X_{\text{norm}}))) \tag{14}$$

$$X_M = \text{DCTConv}\,(X_{\text{scan}}) \odot \text{SiLU}\,(X_{\text{norm}}) + X_{\text{in}} \tag{15}$$

where X_{in} denotes the input features. The addition operation in (15) represents a residual connection enhanced by the activation function. The symbol $\odot$ denotes elementwise multiplication. All feature maps are projected to the same channel dimension to enable residual addition.

To overcome the inherent directionality bias of traditional 1-D scanning, the DCT–Mamba module employs a multidirectional DCT–Mamba Scan mechanism. Unlike standard methods that process sequences in a single order, our mechanism aggregates contextual information along three distinct trajectories—horizontal, vertical, and diagonal—directly on the frequency-domain features extracted by DCTConv. This multipath strategy is essential for addressing the spatial heterogeneity inherent in hyperspectral imagery. It effectively preserves the continuous spatial structure of terrestrial objects, mitigating the loss of long-range geographic dependencies and significantly improving the consistency of spatial modeling in the spectral domain.

2) Transformer Branches: The Transformer component processes spectral patterns through self-attention

$$X_T = \text{DCTConv}\,(\text{DCT-Transformer}\,(\text{DCTConv}\,(X_{\text{norm}}))). \tag{16}$$

D. Feature Enhancement Module: Multiscale Fusion

After extracting global and local features through the DCT–Transformer and DCT–Mamba streams, the feature enhancement module refines these features in the spatio-frequency domain. It ensures that both spatial and spectral dependencies are fully utilized before classification. The feature enhancement module aggregates the outputs from the hybrid module and raw embeddings

$$X_{\text{out}} = \text{Proj}\,(\text{DCTConv}\,(\text{Concat}\,(X_T, X_M))) + X_{\text{norm}}. \tag{17}$$

V. Experiments and Results

We evaluate DTM on three benchmark hyperspectral datasets, which cover diverse land cover types. DTM is tested under different training sample ratios and compared against state-of-the-art models, including CNN-based, Transformer-based, and Mamba-based approaches. Following the common practice in hyperspectral image classification, we employed random sampling to allow fair and direct comparison with a wide range of prior works.

A. Datasets and Preprocessing

We evaluate the DTM architecture on three benchmark HSI datasets.

1) *IPs [30]*: Comprising 145×145 pixels and 200 spectral bands, this dataset includes 16 land cover classes.
2) *Kennedy Space Center [31] (KSC)*: A 512×614 pixel dataset with 176 spectral bands and 13 land cover classes in wetland ecosystems.
3) *Houston [32] (HU)*: This dataset consists of 349×1905 pixels, 144 spectral bands, and 15 land cover classes.

We will specify the exact preprocessing steps applied to each dataset (IPs, KSC, and Houston). This will include as follows.

1) *Spectral Normalization:* Each spectral band was normalized to a zero mean and unit variance: $X\_\text{norm} = (X - \mu)/\sigma$.
2) *Pixel Sampling:* Following common practice in HSI classification, a fixed number of labeled pixels per class were randomly selected to form the training and testing sets. The specific training sample ratio (e.g., 10%) will be stated explicitly for the main comparative experiments.
3) *Patch Extraction:* For each sample pixel, a 3-D spatial–spectral patch of size $P \times P \times C$ was extracted, where P is the spatial patch size and C is the number of spectral bands. The central pixel's label was assigned to the entire patch.

B. Implementation Details

We create a dedicated section that details.

1) *Hardware and Software:* Training platform (Intel Core i7, NVIDIA RTX 3090 Ti), deep learning framework (PyTorch/TensorFlow), and version.
2) *Model Hyperparameters:* DCTConv parameters (number of filters 256, kernel size 3×3), Mamba parameters (state dimension 32, number of layers 3), Transformer parameters (number of heads 8, hidden dimension 128).
3) *Training Strategy:* Detail the optimizer used (e.g., AdamW optimizer), learning rate schedule (0.001, with a cosine annealing schedule), batch size (100), number of epochs (100), loss function (cross-entropy loss), and regularization [weight decay (specify value, e.g., 0.01)]. The input spatial–spectral patch size P is set to 15×15 for all datasets in our experiments, and any data augmentation techniques are applied during training.

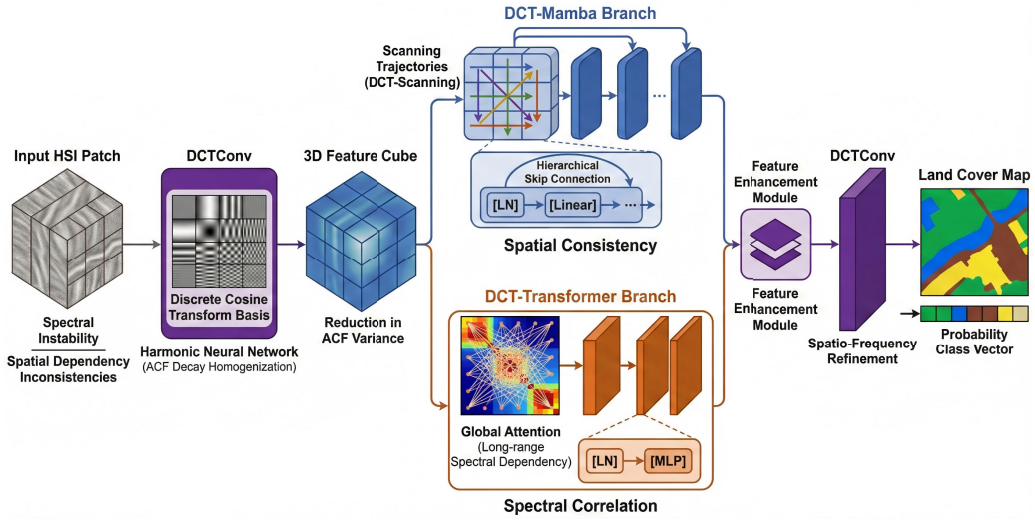

Fig. 2. Overview of the DTM architecture. The input HSI patch undergoes frequency-domain homogenization via the DCT-augmented processing stream (DCTConv). The stabilized features are then processed in parallel: the DCT–Mamba branch captures long-range spatial dependencies via multidirectional SSM, while the DCT–Transformer branch models global spectral correlations via self-attention. Finally, the feature enhancement module fuses the spatial and spectral features for classification. Solid lines indicate the main feature flow; dashed lines represent skip connections.

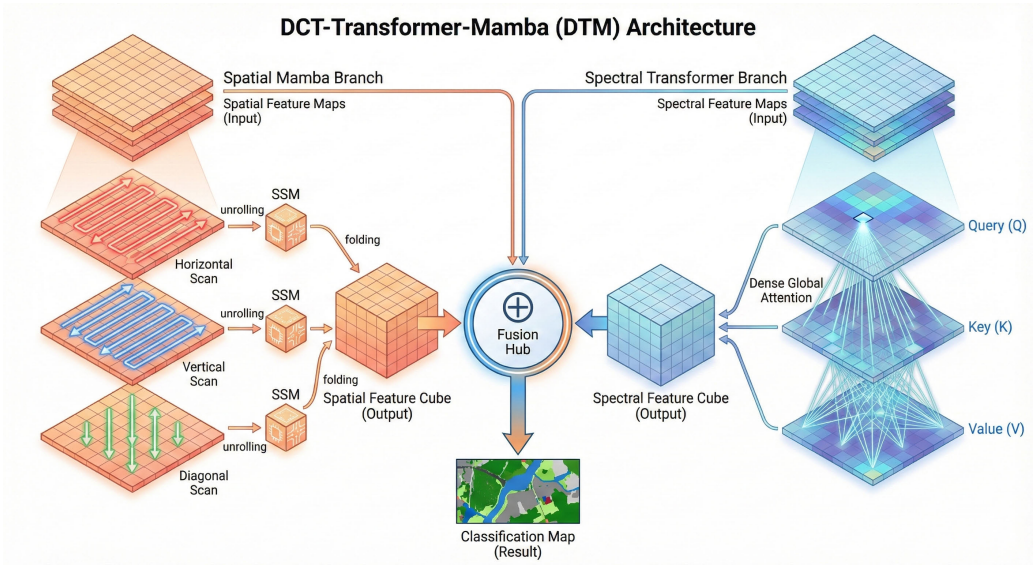

Fig. 3. Schematic of the DTM module integrating a DCT-augmented processing stream with DCT–Transformer and DCT–Mamba components. The figure illustrates the data flow from the spatial feature maps through the multidirectional scanning mechanisms (horizontal, vertical, and diagonal) and the fusion hub.

4) *Evaluation Protocol:* We explicitly state that all experiments were repeated ten times with different random seeds for training/testing splits, and the mean ± standard deviation of performance metrics is reported to ensure statistical robustness.

C. Evaluation Metrics

We evaluate the performance of our model using the following metrics.

1) *Overall Accuracy (OA):* The ratio of correctly classified pixels to the total number of pixels.
2) *Average Accuracy (AA):* The mean of class-wise accuracies, which accounts for class imbalance.
3) *Kappa Coefficient:* A measure of interrater agreement that adjusts for chance agreement.

D. ACF Analysis

In this section, we present an in-depth analysis of the impact of HNNs processing on the ACF of hyperspectral images. The analysis is conducted from two complementary perspectives: 1) mean radial analysis to quantify long-range dependencies over distance, and 2) 2-D spatial coherence validation to visualize directional textures and structural consistency.

1) Mean Radial ACF Analysis (Distance Decay): Instead of treating spatial dimensions as simple 1-D scanning sequences, we computed the mean radial ACF from 2-D patches. This method averages the correlation at each radial distance from

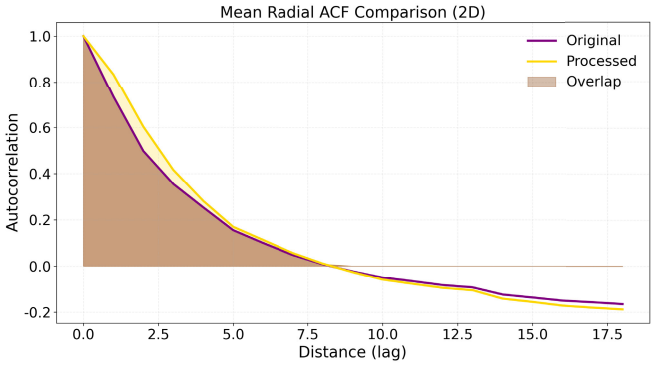

Fig. 4. Mean radial ACF profiles. The plot compares the spatial correlation decay over distance (lag). In the initial lags, the processed curve (yellow) exhibits a significantly slower decay compared to the original curve (purple), demonstrating the enhancement of long-range spatial dependencies. The negative correlations at larger lags reveal the inherent spatial periodicity (texture) revealed after denoising.

TABLE I

ACF DECAY CONSTANTS (τ) FOR THE ORIGINAL AND DCT-AUGMENTED PROCESSED DATA ON THE IPS DATASET

Metric	Original (τ)	Processed (τ)	Change (%)
Mean	2.80	3.18	+13.57%
Variance	0.4832	0.4722	−2.28%

the center, effectively capturing isotropic spatial dependencies regardless of direction.

We calculate the spatial distribution correlation of all pixels within a fixed band to evaluate the long-range dependencies. The τ values are estimated using the exponential decay model $R(k) \sim e^{-k/\tau}$. Fig. 4 illustrates the mean radial ACF profiles. The original curve (purple) drops precipitously near-zero lag, reflecting the interference of high-frequency noise. Conversely, the processed curve (yellow) exhibits a much gentler slope, maintaining higher correlation values over longer distances. This visually confirms that the processing suppresses noise artifacts while preserving long-range spatial dependencies.

Further quantitative analysis, as shown in Fig. 1, confirms these observations: the mean τ increased by **13.57%** (from 2.80 to 3.18), indicating an extended correlation range. Furthermore, the variance of τ was reduced by **2.28%** (from 0.4832 to 0.4722), providing statistical evidence that the spatial dependencies have become more consistent. The variance reduction is calculated as

$$\text{Variance Reduction} = \frac{\text{Var}_{\text{original}} - \text{Var}_{\text{processed}}}{\text{Var}_{\text{original}}} \times 100\%. \quad (18)$$

The results are summarized in Table I.

2) 2-D Spatial Coherence Validation (Directional Structure): While the mean radial ACF summarizes the global decay trend, it averages out directional information. To further validate the improvement in spatial structure, we conducted a 2-D ACF analysis.

Fig. 5 presents the 2-D autocorrelation maps. The difference map quantifies improvements, where blue regions indicate noise suppression and red areas show enhanced consistency. The original data exhibits irregular decay patterns with localized discontinuities due to noise. After DCT-augmented processing, the spatial coherence becomes more homogeneous and smoother, as shown by the 2-D ACF map. This comparison highlights that our method not only extends the correlation distance (as shown in Fig. 4) but also restores the **directional texture patterns** and spatial anisotropy that are critical for identifying complex land covers.

E. Comparative Analysis

We compare the DTM framework against state-of-the-art methods in HSI classification, including CNN-based methods [33] (2D-CNN [34], 3D-CNN [35], and HybridSN [36]), Transformer-based methods (ViT [21], HiT [37], MorphF [38], SSFTT [39]), MambaHSI [24], and Mamba-based methods (MiM [17]). We evaluate classification accuracy and Kappa coefficient for each method across the IPs, KSC, and Houston datasets, as shown in Tables II–IV. However, we visualized the three data separately, as shown in the Figs. 6–8.

The DTM framework demonstrates superior performance across all three datasets, achieving overall accuracies of **95.72%** on IP, **98.76%** on KSC, and **98.50%** on HU. In challenging categories such as "Stone–Steel–Towers" on IP and "Scrub" on KSC, DTM outperforms other models, including Mamba and Transformer-based methods, by significant margins. DTM's ability to handle spectral variability and high spectral similarity is particularly evident in classes with fine-grained differences, such as "Soybean-mintill" and "Healthy Grass," as shown in Tables II and IV.

DTM achieves superior performance across all datasets, handling spectral variability, high spectral similarity, and heterogeneous backgrounds with remarkable accuracy. Notably, DTM achieves overall accuracies of **95.72%** on IP, **98.76%** on KSC, and **98.50%** on HU, with corresponding Kappa coefficients of **95.11%**, **98.62%**, and **98.37%**. The model exhibits rapid convergence, reaching near-zero loss by the **5th epoch**, significantly faster than its competitors.

In challenging categories, such as "Stone–Steel–Towers" on IP, "Scrub" on KSC, and "Residential" on HU, DTM outperforms the Mamba model by up to **56.69%** and Transformer models by up to **36.91%**. In classes with high spectral similarity, such as "Soybean-mintill" on IP, "Cabbage Palm Hammock" on KSC, and "Healthy Grass" on HU, DTM consistently surpasses Transformer models by margins of up to **79.47%** and Mamba models by up to **5.69%**.

Moreover, DTM excels in both natural and controlled environments, achieving better accuracies in consistent spectral classes like "Salt Marsh" on KSC and "Soil" on HU, and excellent performance in controlled environments such as "Tennis Court" on HU. Even in small sample classes like "Alfalfa" on IP, DTM achieves **92.99%** accuracy, demonstrating strong generalization capabilities.

F. T-SNE Visualization and Patch Size Analysis

To further evaluate DTM's performance, we use t-distributed stochastic neighbor embedding (t-SNE) for visualization, as depicted in Fig. 9. The results indicate that DTM achieves more compact and well-separated clusters

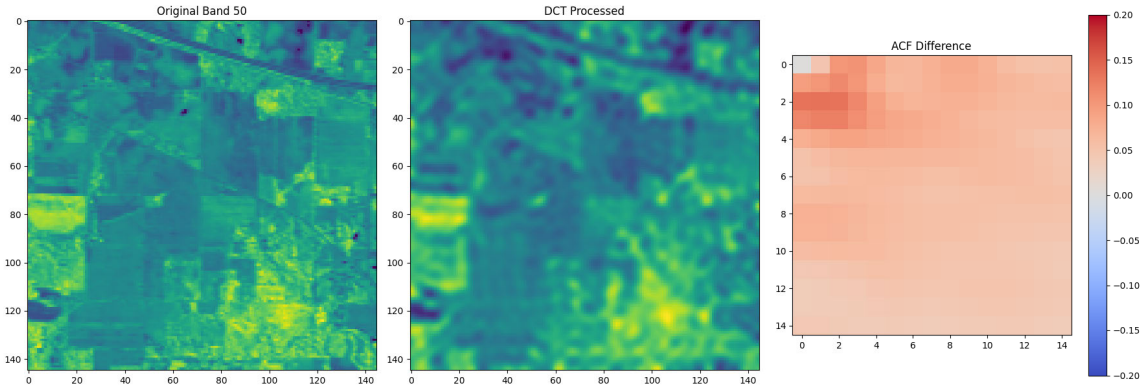

Fig. 5. Comparison of 2-D ACF before and after DCT-augmented processing. (a) Original band 50. (b) DCT processed data showing enhanced spatial coherence. (c) ACF difference map highlighting the improvements in spatial consistency.

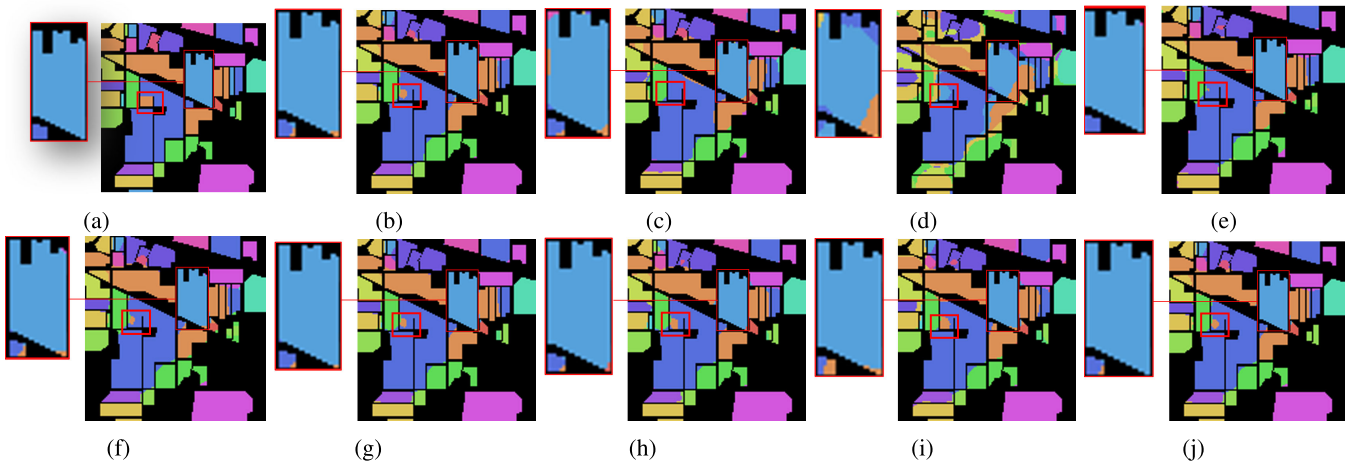

Fig. 6. Classification results on the IPs dataset using various methods with 10% training samples. (a) Ground truth. (b) 2D-CNN. (c) 3D-CNN. (d) HybridSN. (e) ViT. (f) HiT. (g) MorphF. (h) SSFTT. (i) MiM. (j) Ours.

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART TRANSFORMERS AND MAMBA ON IPS DATASET (10% TRAINING SAMPLES)

Class	2D-CNN	3D-CNN	HybridSN	ViT	HiT	MorphFormer	SSFTT	MiM	MambaHSI	DTM
Alfalfa	92.82 ± 4.80	70.82 ± 18.73	34.67 ± 31.07	76.55 ± 8.83	88.77 ± 3.59	79.58 ± 15.39	89.95 ± 8.31	76.20 ± 8.43	71.71 ± 17.49	**92.99 ± 2.10**
Corn-notill	93.81 ± 1.97	89.33 ± 1.03	88.50 ± 4.59	94.10 ± 0.36	94.45 ± 0.61	93.08 ± 1.67	92.40 ± 2.46	92.49 ± 0.78	92.45 ± 3.44	**95.23 ± 0.56**
Corn-mintill	92.19 ± 1.77	87.44 ± 2.55	81.40 ± 10.60	93.56 ± 0.82	**94.11 ± 0.66**	90.15 ± 2.60	88.62 ± 4.68	89.03 ± 2.22	92.90 ± 3.03	92.68 ± 0.43
Corn	97.94 ± 1.50	94.78 ± 1.36	83.47 ± 10.37	99.58 ± 0.31	99.73 ± 0.21	92.96 ± 3.49	96.23 ± 2.09	93.06 ± 1.83	85.07 ± 3.62	**99.77 ± 0.18**
Grass-pasture	93.09 ± 3.32	92.88 ± 0.97	84.74 ± 7.30	91.34 ± 0.93	92.75 ± 0.65	**95.13 ± 1.39**	93.71 ± 2.60	93.20 ± 1.69	93.47 ± 1.37	94.78 ± 0.41
Grass-trees	95.65 ± 2.97	94.41 ± 1.15	82.42 ± 12.40	89.85 ± 1.38	93.12 ± 0.92	95.80 ± 0.93	94.43 ± 1.89	96.01 ± 0.56	95.25 ± 2.26	**97.28 ± 0.48**
Grass-pasture-mowed	7.94 ± 18.29	0.00 ± 0.00	1.21 ± 3.64	0.00 ± 0.00	0.00 ± 0.00	**66.25 ± 28.01**	56.16 ± 35.15	32.43 ± 15.11	81.60 ± 22.57	34.50 ± 17.16
Hay-windrowed	99.69 ± 0.55	99.09 ± 0.61	92.37 ± 8.37	99.83 ± 0.16	99.59 ± 0.20	99.81 ± 0.35	99.23 ± 1.11	99.92 ± 0.11	98.56 ± 0.77	**99.97 ± 0.05**
Oats	**73.30 ± 29.09**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	9.19 ± 27.57	38.88 ± 32.47	53.87 ± 20.58	45.56 ± 12.86	38.69 ± 34.02
Soybean-notill	87.78 ± 1.60	83.76 ± 1.43	82.54 ± 5.97	**89.52 ± 0.72**	89.44 ± 0.27	89.35 ± 2.41	87.84 ± 1.93	86.27 ± 1.69	94.74 ± 1.04	88.83 ± 0.26
Soybean-mintill	96.26 ± 1.24	94.33 ± 0.63	93.02 ± 1.91	96.58 ± 0.17	96.65 ± 0.12	96.71 ± 0.59	96.08 ± 1.03	96.27 ± 0.53	96.90 ± 1.08	**97.25 ± 0.17**
Soybean-clean	91.80 ± 2.21	89.17 ± 2.54	81.26 ± 5.65	91.69 ± 1.24	93.36 ± 0.37	88.93 ± 4.23	87.39 ± 6.02	84.99 ± 3.09	88.20 ± 3.45	**94.71 ± 0.21**
Wheat	98.12 ± 1.32	86.12 ± 11.32	47.40 ± 41.31	97.28 ± 1.21	97.59 ± 0.71	92.52 ± 7.20	93.27 ± 4.93	91.09 ± 4.65	90.70 ± 6.48	**98.58 ± 0.78**
Woods	98.28 ± 2.42	97.96 ± 0.67	97.28 ± 0.86	98.20 ± 0.25	98.43 ± 0.22	98.87 ± 0.90	98.88 ± 0.60	98.04 ± 0.63	98.00 ± 1.37	**99.19 ± 0.13**
Buildings-Grass-Trees-Drives	97.82 ± 1.46	92.51 ± 1.86	74.80 ± 20.55	98.05 ± 0.69	98.46 ± 0.70	95.08 ± 1.59	94.88 ± 4.08	92.95 ± 1.68	89.80 ± 3.54	**98.88 ± 0.22**
Stone-Steel-Towers	52.74 ± 21.39	51.18 ± 10.42	15.74 ± 23.97	34.79 ± 20.21	67.40 ± 2.92	55.80 ± 32.60	34.81 ± 32.68	15.01 ± 21.53	85.71 ± 2.71	**71.70 ± 2.20**
Accuracy (%)	94.48 ± 1.41	91.65 ± 0.62	86.81 ± 5.68	94.18 ± 0.29	94.91 ± 0.18	94.14 ± 0.72	93.36 ± 1.43	92.81 ± 0.65	94.16 ± 1.59	**95.72 ± 0.16**
Kappa (%)	93.69 ± 1.61	90.45 ± 0.71	84.87 ± 6.53	93.35 ± 0.33	94.19 ± 0.21	93.30 ± 0.83	92.42 ± 1.63	91.78 ± 0.74	93.34 ± 1.81	**95.11 ± 0.18**

compared to other methods, highlighting its ability to reduce misclassification and enhance feature cohesion.

In addition, we examine the impact of patch size on classification accuracy. Smaller patch sizes (e.g., 9×9) yield the best performance, with accuracy rates of **99.13%**, **99.65%**, and **99.39%** on the IPs, as shown in Table V, KSC, and Houston datasets, respectively. Larger patch sizes lead to a decline in accuracy, but DTM still outperforms other methods even with larger patch sizes.

1) Impact of Training Sample Ratios on Classification Accuracy: Table VI shows the performance of DTM and baseline Mamba model (base-Mamba) across different training sample ratios on the Houston2013, IPs, and KSC datasets. The results indicate that DTM maintains high accuracy even when trained with a smaller proportion of samples. For instance, on the Houston2013 dataset, the model achieves **98.50%** accuracy using only 10% of the data for training, and this performance remains stable as the training sample ratio increases.

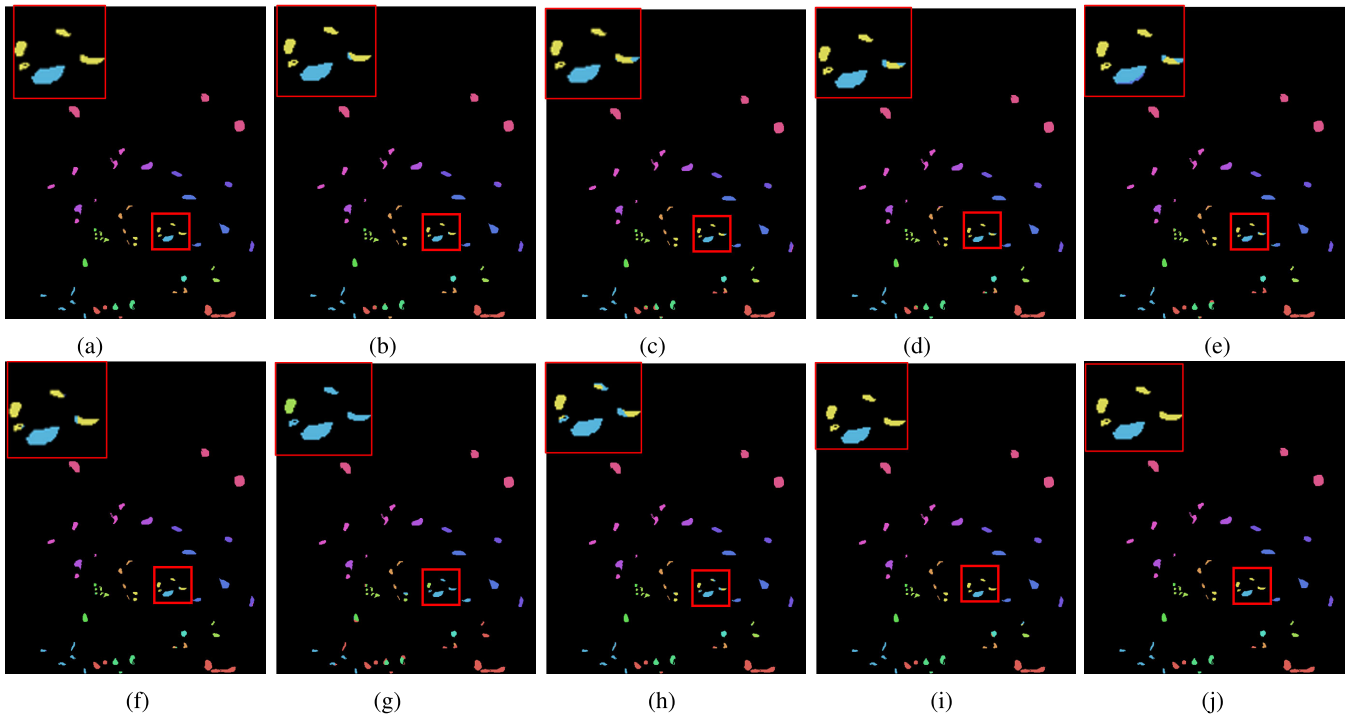

Fig. 7. Classification results on the KSC dataset using various methods with 10% training samples. (a) Ground truth. (b) 2D-CNN. (c) 3D-CNN. (d) HybridSN. (e) ViT. (f) HiT. (g) MorphF. (h) SSFTT. (i) MiM. (j) Ours.

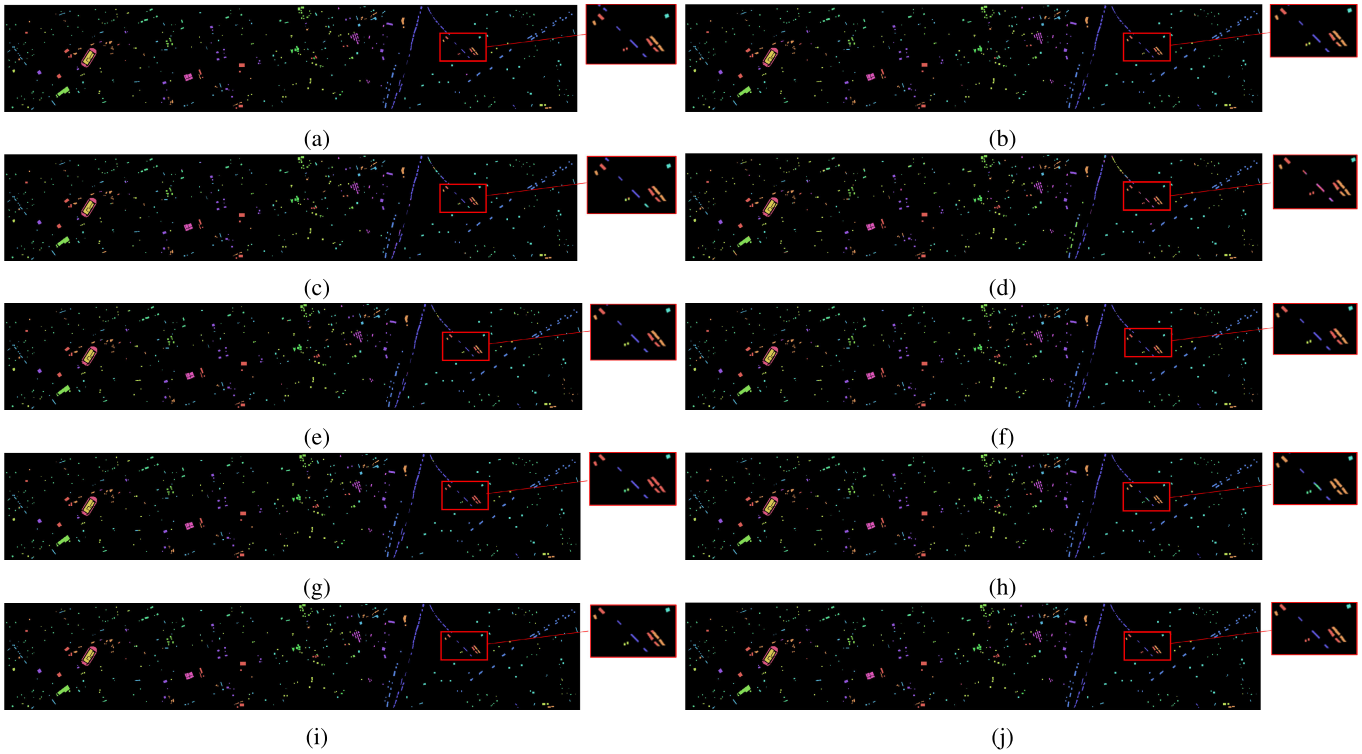

Fig. 8. Classification results on the Houston2013 scene dataset using various methods with 10% training samples. (a) Ground truth. (b) 2D-CNN. (c) 3D-CNN. (d) HybridSN. (e) ViT. (f) HiT. (g) MorphF. (h) SSFTT. (i) MiM. (j) Ours.

Similarly, the model performs consistently well on the KSC dataset, with only minor variations as the training sample ratio increases. These results suggest that DTM has strong generalization capability and can perform effectively even with limited labeled data.

G. Ablation Study of Key Modules

To comprehensively evaluate the contributions of the core components within the DTM architecture, we conducted an ablation study based on its three primary modules: 1) the DCT-augmented processing stream (DCTConv) for spatial

TABLE III

COMPARISON WITH THE STATE-OF-THE-ART TRANSFORMERS AND MAMBA ON KSC DATASET (10% TRAINING SAMPLES)

Class	2D-CNN	3D-CNN	HybridSN	ViT	HiT	MorphFormer	SSFTT	MiM	MambaHSI	DTM
Scrub	97.38 ± 1.15	95.96 ± 1.60	95.89 ± 1.24	91.62 ± 1.37	86.13 ± 0.08	80.24 ± 13.43	92.20 ± 10.12	98.32 ± 1.10	95.42 ± 1.17	**98.88 ± 0.58**
Willow swamp	95.16 ± 3.99	87.20 ± 3.13	87.77 ± 3.31	74.05 ± 2.53	**98.69 ± 1.02**	57.59 ± 28.29	88.83 ± 7.55	93.30 ± 3.92	55.21 ± 21.04	98.67 ± 1.47
Cabbage palm hammock	97.78 ± 2.27	91.63 ± 3.69	94.18 ± 3.20	85.03 ± 6.24	99.17 ± 0.70	20.20 ± 33.30	49.84 ± 38.52	93.98 ± 4.42	79.57 ± 9.26	**99.67 ± 0.75**
Cabbage palm/oak hammock	92.82 ± 4.59	83.38 ± 4.82	89.29 ± 3.94	93.50 ± 4.18	**98.36 ± 1.43**	52.04 ± 21.72	82.66 ± 16.70	74.07 ± 4.05	71.10 ± 18.25	94.67 ± 4.34
Slash pine	89.94 ± 6.01	78.78 ± 9.74	78.63 ± 10.38	94.89 ± 3.09	98.83 ± 1.29	70.67 ± 12.55	95.07 ± 5.10	21.99 ± 28.44	83.79 ± 12.94	**98.84 ± 1.71**
Oak/broadleaf hammock	94.68 ± 2.11	93.27 ± 3.55	90.85 ± 3.37	93.21 ± 2.43	**99.74 ± 0.33**	41.96 ± 39.57	90.31 ± 16.47	93.72 ± 5.08	83.25 ± 22.74	98.37 ± 2.13
Hardwood swamp	98.79 ± 2.30	99.32 ± 1.10	99.05 ± 0.94	**99.50 ± 1.50**	99.02 ± 1.51	25.76 ± 40.30	66.07 ± 43.75	91.61 ± 6.94	81.81 ± 28.56	99.38 ± 1.88
Graminoid marsh	94.66 ± 12.39	97.62 ± 1.44	96.55 ± 1.85	89.51 ± 2.20	96.78 ± 0.16	67.25 ± 15.14	85.22 ± 9.82	**99.62 ± 0.85**	88.17 ± 8.81	99.59 ± 1.00
Spartina marsh	97.49 ± 6.66	99.70 ± 0.39	99.40 ± 0.95	95.99 ± 1.24	99.99 ± 0.03	73.41 ± 30.31	81.59 ± 32.05	99.89 ± 0.24	91.79 ± 4.81	**99.99 ± 0.03**
Cattail marsh	97.17 ± 1.49	97.37 ± 2.25	98.36 ± 1.84	97.75 ± 1.30	99.03 ± 0.53	89.24 ± 21.11	**100.00 ± 0.00**	**100.00 ± 0.00**	96.76 ± 4.19	94.63 ± 2.83
Salt marsh	99.65 ± 0.61	**99.98 ± 0.07**	98.06 ± 3.88	**100.00 ± 0.00**	**100.00 ± 0.00**	97.67 ± 2.44	98.28 ± 5.16	**100.00 ± 0.00**	98.54 ± 1.37	**100.00 ± 0.00**
Mud flats	98.28 ± 1.01	97.85 ± 1.77	99.08 ± 0.74	99.65 ± 0.30	99.17 ± 0.16	95.03 ± 6.03	**100.00 ± 0.00**	**100.00 ± 0.00**	99.98 ± 0.07	98.26 ± 2.27
Water	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	99.86 ± 0.08	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	100.00 ± 0.00	**100.00 ± 0.00**
Accuracy (%)	97.31 ± 1.54	95.99 ± 1.08	96.28 ± 0.69	94.77 ± 0.76	95.51 ± 0.17	80.36 ± 10.22	91.99 ± 7.72	95.80 ± 0.77	91.07 ± 4.03	**98.76 ± 0.35**
Kappa (%)	97.01 ± 1.72	95.54 ± 1.21	95.86 ± 0.77	94.16 ± 0.85	95.03 ± 0.19	77.82 ± 11.72	91.02 ± 8.70	95.32 ± 0.86	90.02 ± 4.53	**98.62 ± 0.39**

TABLE IV

COMPARISON WITH THE STATE-OF-THE-ART TRANSFORMERS AND MAMBA ON HOUSTON2013 DATASET (10% TRAINING SAMPLES)

Class	2D-CNN	3D-CNN	HybridSN	ViT	HiT	MorphFormer	SSFTT	MiM	MambaHSI	DTM
unclassified	95.45 ± 1.95	93.10 ± 1.01	92.46 ± 2.00	93.45 ± 0.87	94.63 ± 0.58	93.78 ± 2.03	90.18 ± 6.29	93.36 ± 1.08	96.77 ± 1.94	**96.98 ± 0.92**
Healthy Grass	96.45 ± 1.80	90.56 ± 2.94	88.22 ± 6.06	87.16 ± 2.54	90.15 ± 2.28	94.20 ± 2.57	92.70 ± 4.37	95.09 ± 1.62	96.38 ± 1.92	**97.94 ± 0.78**
Stressed grass	99.32 ± 0.50	97.35 ± 2.62	97.84 ± 1.81	97.29 ± 0.96	98.93 ± 0.24	99.18 ± 0.41	**99.52 ± 0.41**	99.03 ± 0.31	98.79 ± 0.45	99.38 ± 0.13
Synthetic Grass	92.42 ± 1.58	89.78 ± 1.98	81.79 ± 5.17	85.45 ± 2.05	87.57 ± 1.55	**97.76 ± 1.03**	93.53 ± 1.76	91.13 ± 2.44	95.23 ± 2.03	96.69 ± 0.98
Soil	99.57 ± 0.17	97.53 ± 2.08	97.47 ± 2.99	99.56 ± 0.12	99.66 ± 0.16	99.74 ± 0.20	99.33 ± 0.52	99.90 ± 0.08	99.94 ± 0.11	**99.94 ± 0.07**
Water	95.02 ± 3.99	86.73 ± 2.30	93.09 ± 2.89	90.34 ± 2.07	88.05 ± 1.85	94.86 ± 1.36	**93.44 ± 2.33**	91.60 ± 2.97	90.31 ± 3.44	93.20 ± 2.13
Residential	93.96 ± 1.75	86.51 ± 2.67	86.07 ± 8.24	89.16 ± 1.17	90.17 ± 1.22	97.10 ± 1.80	95.40 ± 2.45	90.88 ± 3.91	98.31 ± 1.02	**98.25 ± 0.63**
Commercial	96.59 ± 2.21	89.12 ± 1.31	92.09 ± 6.29	96.86 ± 0.72	96.95 ± 0.97	98.21 ± 1.26	96.59 ± 1.37	**99.39 ± 0.46**	98.51 ± 0.85	99.25 ± 0.47
Road	94.59 ± 1.57	84.52 ± 2.49	76.40 ± 15.88	90.09 ± 1.34	89.77 ± 1.25	96.95 ± 2.16	95.06 ± 1.45	93.98 ± 2.27	98.19 ± 0.89	**96.96 ± 0.91**
Highway	97.30 ± 2.79	94.23 ± 2.12	94.67 ± 5.05	99.17 ± 0.26	97.88 ± 0.53	99.20 ± 0.44	99.22 ± 1.19	**99.73 ± 0.32**	99.92 ± 0.24	99.60 ± 0.36
Railway	99.41 ± 0.47	87.33 ± 2.33	82.52 ± 18.39	98.29 ± 1.76	99.50 ± 0.76	**99.85 ± 0.20**	99.56 ± 0.66	98.10 ± 2.15	99.89 ± 0.24	99.84 ± 0.28
Parking Lot 1	97.78 ± 2.78	95.51 ± 1.66	96.37 ± 2.01	98.04 ± 0.53	98.54 ± 0.52	99.45 ± 0.37	98.21 ± 1.04	99.24 ± 0.42	98.93 ± 0.60	**99.85 ± 0.24**
Parking Lot 2	96.27 ± 1.40	88.31 ± 3.55	87.86 ± 9.65	94.16 ± 1.52	94.78 ± 1.91	97.38 ± 1.57	97.69 ± 1.07	96.77 ± 1.47	97.42 ± 2.08	**97.81 ± 1.17**
Tennis Court	99.92 ± 0.21	98.84 ± 1.18	96.61 ± 4.72	99.85 ± 0.28	99.95 ± 0.09	99.87 ± 0.24	99.51 ± 0.62	99.78 ± 0.13	100.00 ± 0.00	**100.00 ± 0.00**
Running Track	98.92 ± 0.61	95.49 ± 3.84	94.23 ± 8.46	96.77 ± 1.00	98.75 ± 0.35	98.96 ± 0.36	**99.15 ± 0.49**	98.81 ± 0.47	100.00 ± 0.00	99.00 ± 0.25
Accuracy (%)	96.66 ± 0.86	91.36 ± 1.13	90.09 ± 4.70	94.09 ± 0.62	94.87 ± 0.48	97.75 ± 0.72	96.37 ± 1.17	96.32 ± 0.85	97.86 ± 0.34	**98.50 ± 0.37**
Kappa (%)	96.39 ± 0.93	90.65 ± 1.22	89.29 ± 5.07	93.61 ± 0.68	94.45 ± 0.52	97.57 ± 0.78	96.08 ± 1.26	96.02 ± 0.92	98.01 ± 0.37	**98.37 ± 0.40**

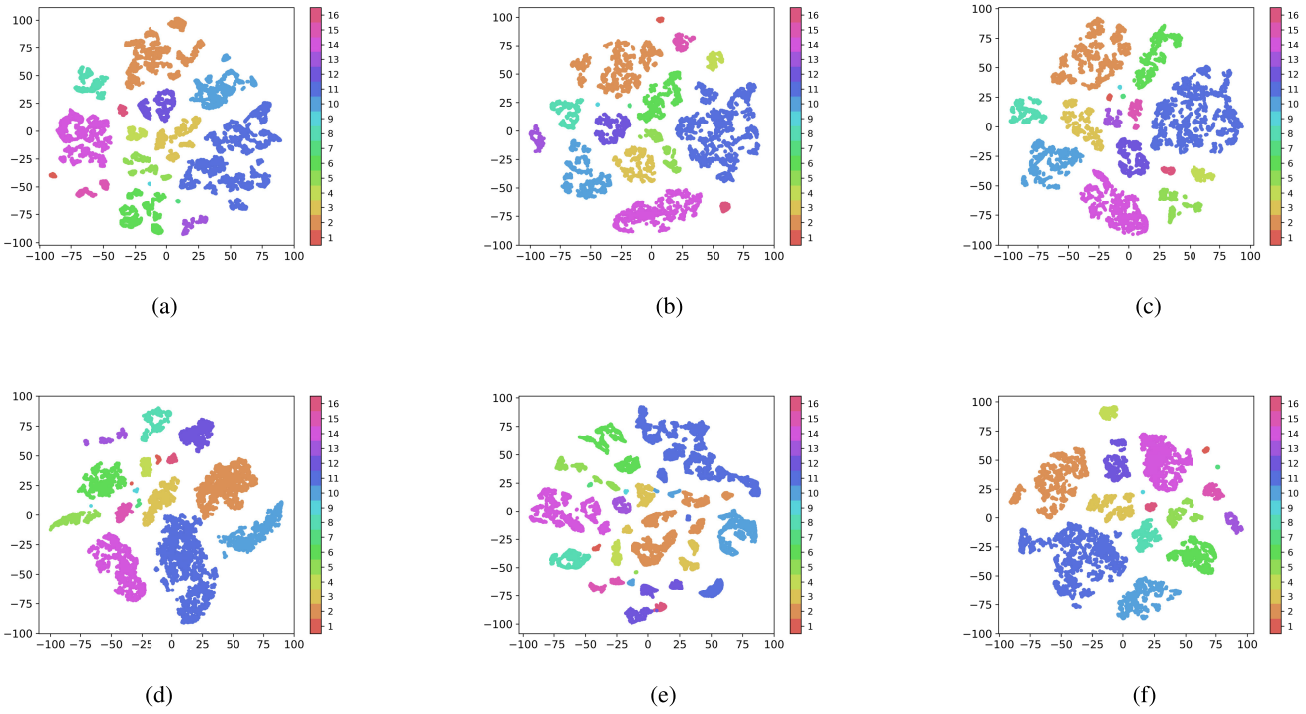

Fig. 9. T-SNE visualizations for different algorithms on the IPs dataset. (a) 2D-CNN. (b) ViT. (c) HiT. (d) HybridSN. (e) MiM. (f) Ours.

consistency enhancement; 2) the hybrid feature extraction module for spectral–spatial feature modeling (two branches); and 3) the feature enhancement module for improving representational quality. The results, presented in Table VII, provide

TABLE V
CLASSIFICATION ACCURACY WITH DIFFERENT PATCH SIZES ACROSS MULTIPLE DATASETS

Dataset	Patch Size			
	9x9	11x11	13x13	15x15
Indian Pines	99.13 ± 0.15	98.17 ± 0.06	97.03 ± 0.11	95.72 ± 0.16
KSC	99.65 ± 0.19	99.53 ± 0.25	99.13 ± 0.44	98.76 ± 0.35
Houston	99.39 ± 0.22	99.22 ± 0.11	98.64 ± 0.39	98.50 ± 0.37

TABLE VI
PERFORMANCE COMPARISON WITH DIFFERENT TRAINING SAMPLE RATIOS ON IPS DATASET

Training Samples	0.1	0.3	0.5	0.7
Houston-base	95.70 ± 0.37	98.19 ± 0.26	98.39 ± 0.33	98.74 ± 0.46
Indian Pines-base	92.50 ± 0.16	94.03 ± 0.11	95.39 ± 0.34	95.25 ± 0.38
KSC-base	91.80 ± 0.35	99.24 ± 0.21	98.96 ± 0.27	99.12 ± 0.22
Houston-our	98.50 ± 0.37	98.74 ± 0.20	98.75 ± 0.29	98.85 ± 0.24
Indian Pines-our	95.72 ± 0.16	95.58 ± 0.11	95.58 ± 0.15	95.65 ± 0.12
KSC-our	98.76 ± 0.35	98.77 ± 0.36	98.80 ± 0.24	98.92 ± 0.29

TABLE VII
ABLATION STUDY OF KEY MODULES (ACCURACY IMPROVEMENTS IN %)

Method	**Indian Pines**	**KSC**	**Houston**
Baseline (only mamba branches without DCTConv)	92.50	91.80	95.70
+ DCTConv	+2.98	+4.95	+2.26
+ Transformer branches without DCT-Transformer	+1.80	-1.46	-0.80
+ Transformer branches with DCT-Transformer	+2.44	+5.98	-0.44
+ Feature Enhancement Module	+2.09	+6.82	+1.92
DTM block without Mamba	**-0.39**	**+2.80**	**-0.48**
DTM block without DCT-Mamba Scan	**+0.46**	**+3.06**	**-0.44**
DTM block without DCT	**+0.70**	**+3.78**	**-0.57**
Complete DTM block	**+3.22**	**+6.96**	**+2.80**

insights into how each module contributes to overall model performance.

1) DCT-Augmented Processing Stream: Enhancing Spatial Consistency: One of the fundamental challenges in hyperspectral image (HSI) classification is the presence of spatial dependency inconsistencies, often observed as rapid autocorrelation decay in heterogeneous regions. Through the ACF analysis, we observed that raw HSI data exhibit irregular long-range dependencies, complicating spatial feature modeling.

To address this, we employed a DCT-augmented processing stream that utilizes DCT and HNNs to decompose spatial patterns into frequency components. This frequency-domain processing suppresses high-frequency noise and enhances the spatial consistency of the data. Notably, ACF analysis was not used to guide the DCT processing but served as a diagnostic tool to evaluate the spatial dependency structure before and after processing.

The experimental results demonstrated that after applying the DCT-augmented processing stream, the ACF decay constant (τ) increased by 13.57%, indicating improved long-range dependency modeling. These improvements led to classification accuracy gains of 2.98% on the IPs dataset, 4.95% on the KSC dataset, and 2.26% on the Houston dataset.

2) Hybrid Feature Extraction Module: Bridging Spatial and Spectral Patterns: The hybrid feature extraction module combines the strengths of Mamba and Transformer architectures to bridge spatial and spectral feature extraction. Mamba excels at capturing long-range dependencies but struggles with anisotropic spatial structures when applied independently. As shown in Table VII, introducing the Mamba component (*with Mamba*) improved performance by 1.80% in the IPs dataset but caused a decrease of 1.46% in KSC and 0.80% in Houston due to unresolved spatial irregularities.

Integrating the Transformer module into this hybrid feature extraction design (*mamba-T*) mitigated these issues by providing global spatial context, resulting in a performance increase of 2.44% in IPs and 5.98% in KSC. The module's effectiveness is attributed to the Transformer's capacity to capture spatial patterns across varying distances, while Mamba focuses on sequence-level interactions.

3) Feature Enhancement Module: Refining Discriminative Power: The feature enhancement module is designed to refine both spatial and spectral features by integrating outputs from the DCT-augmented processing stream and the hybrid feature extraction module. It applies a multiscale aggregation strategy to enhance class separability, particularly for spectrally similar classes.

As depicted in Table VII, the addition of this module (*mamba-Stem*) led to performance improvements of 2.09% on IPs, 6.82% on KSC, and 1.92% on Houston. This suggests that the feature enhancement module plays a critical role in ensuring that the frequency-augmented features are properly aligned and discriminative in the classification process.

4) Performance of the Complete DTM Model: The final configuration, referred to as the DTM framework, integrates all three modules into a unified pipeline. As shown in Table VII, the complete DTM architecture achieved the highest

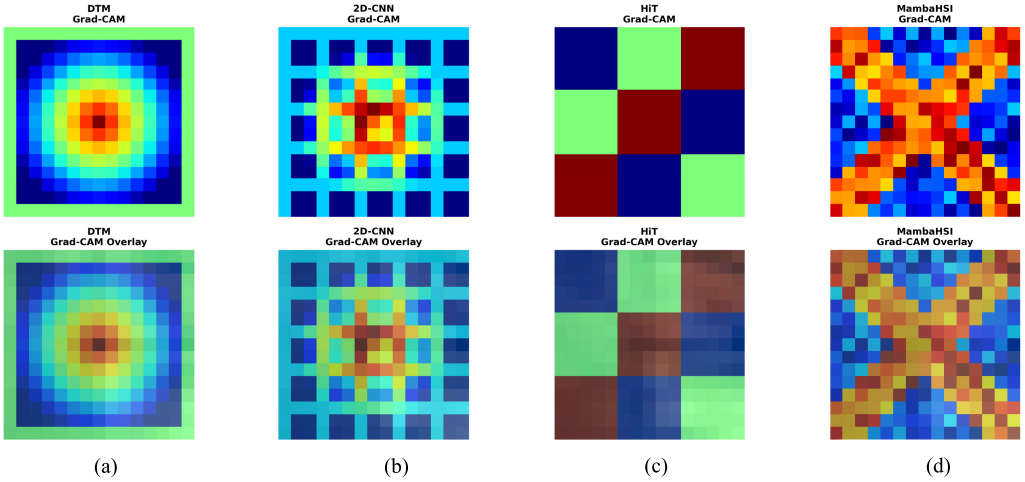

Fig. 10. Class activation map visualization for Stone–Steel–Towers class. (a) DTM attention map, (b) 2D-CNN attention map, (c) HiT attention map, and (d) MambaHSI attention map. DTM demonstrates a focused and accurate attention on the target class, while other models show dispersed or incorrect attention patterns.

performance across all datasets, with gains of 3.22%, 6.96%, and 2.80% for IPs, KSC, and Houston, respectively. These results confirm that the combined effect of spatial consistency enhancement, hybrid feature extraction, and multiscale feature enhancement contributes significantly to improved hyperspectral image classification.

This ablation study demonstrates that the DCT-augmented processing stream effectively increases ACF variance and enhances spatial consistency, while the hybrid feature extraction module supports robust long-range dependency modeling. The physically grounded frequency-domain modeling strategy improves structural consistency of spatial representations, which contributes to more stable and reliable learning behavior in hyperspectral image analysis. The feature enhancement module further refines these representations, leading to a high-performance DTM model that consistently outperforms existing methods in hyperspectral image classification tasks.

In addition, we performed ablation experiments by 1) removing the SSM module (DTM block without Mamba), and 2) substituting the DCT–Mamba scan with a one-way scan. The results clearly highlight the essential contributions of both Mamba and DCT–Mamba scanning to the effectiveness of DTM.

5) Class Activation Map Visualization: The qualitative diagnosis in Fig. 10 offers a compelling visual explanation for the quantitative performance gaps. For the challenging Stone–Steel–Towers class, baseline models exhibit flawed attention mechanisms: 2D-CNN's attention is local but unfocused; HiT's is global but imprecise; MambaHSI's is inconsistent due to raw spatial noise. These flawed attentions directly translate to incorrect or overconfident predictions. Conversely, DTM, benefiting from the spatially homogenized input, generates a sharply focused activation map that pinpoints the discriminative target. This demonstrates that our proposed DCTConv module not only improves a metric (ACF variance) but fundamentally refines the model's interpretive focus, leading to more reliable and accurate pixel-level decisions.

6) Feature Flow Visualization: To elucidate the spatial decision logic of each model, we apply Grad-CAM to a challenging sample (see Fig. 11). The resulting attribution maps reveal distinct patterns: a localized, grid-bound focus for 2D-CNN; a fragmented, blockwise pattern for HiT; and a broad, diffuse activation for MambaHSI, reflecting its long-range but potentially noisy scanning.

In contrast, DTM exhibits a sharply concentrated, radially symmetric activation. We argue this focused pattern is the direct visual evidence of our core innovation—the DCT-augmented processing stream (DCTConv). By homogenizing spatial inconsistencies in the frequency domain, DCTConv acts as an adaptive filter, suppressing high-frequency noise and irrelevant contextual distractions. This produces a denoised and spatially regularized feature representation where discriminative local structures are significantly enhanced.

Thus, the subsequent Mamba and Transformer branches operate on a high signal-to-noise input. The Mamba branch can precisely attend to these salient local cues without being diverted by spurious long-range correlations (as seen in MambaHSI's diffuse map), while the Transformer branch complements this with global spectral context. This efficient division of labor—spatial regularization by DCTConv followed by precise feature extraction—explains DTM's robust performance. The interpretable, focused attribution aligns with and provides a visual rationale for DTM's superior accuracy on complex classes (e.g., Stone–Steel–Towers), demonstrating its ability to base decisions on clear, discriminative evidence.

VI. Discussion

A. Methodological Insights

The DTM architecture introduced a novel approach to hyperspectral image (HSI) classification by addressing spatial dependency inconsistencies through ACF analysis. Unlike traditional methods that primarily adjust model architectures to infer hidden dependencies from irregular spatial patterns,

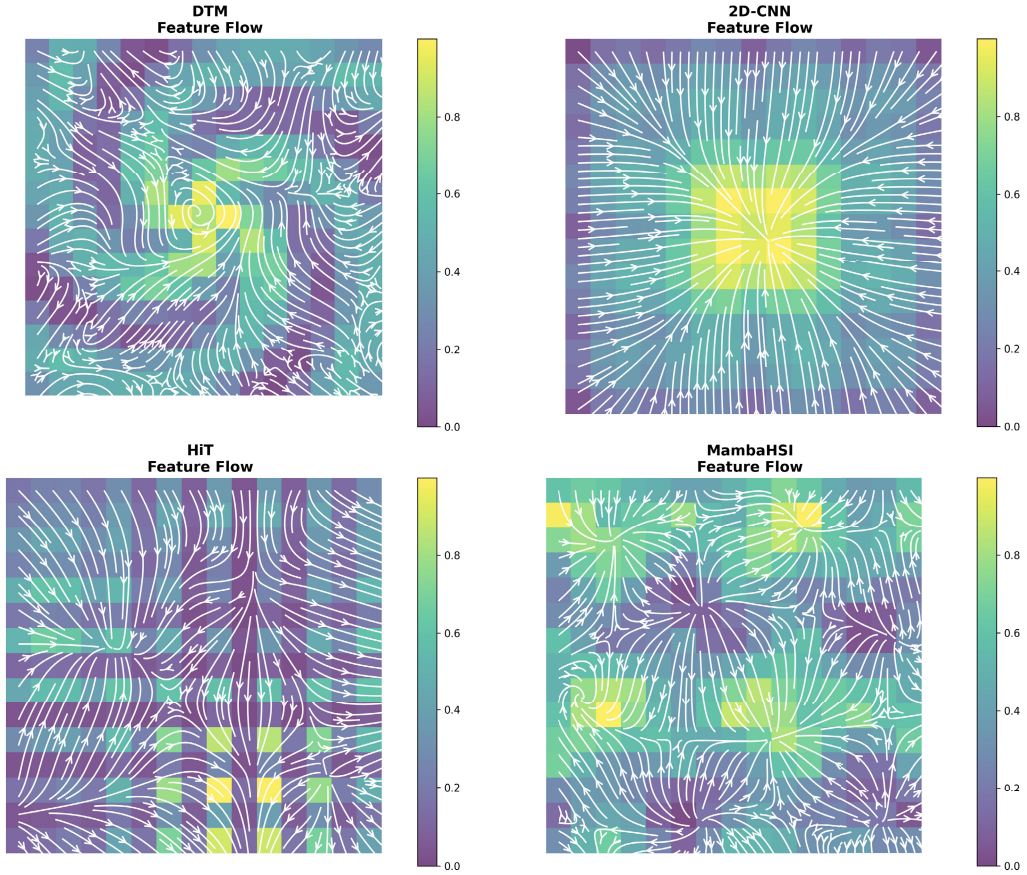

Fig. 11. Feature flow visualization with different methods, including DTM, 2D-CNN, HiT, and MambaHSI.

we proposed a new perspective: transforming unbalanced spatial data into a more ordered form to enhance long-range dependency modeling. The core innovation of DTM lies in the application of DCT-augmented processing combined with HNNs. This approach homogenized spatial autocorrelation by extending the spatial correlation range by **13.57%**, thereby improving the spatial consistency necessary for robust sequential modeling.

By enhancing spatial consistency, DTM improved Mamba's ability to model long-range dependencies, overcoming its inherent 1-D sequence modeling bias. The integration of bidirectional scanning mechanisms further contributed to capturing complex spatial structures across heterogeneous regions, while the hybrid architecture—consisting of DCT, Transformer, and Mamba components—achieved effective spectral–spatial disentanglement.

B. Performance Analysis

The improved spatial consistency and long-range dependency modeling directly translated into better classification performance across various datasets. The experimental results showed that DTM achieved a state-of-the-art accuracy of **98.76%** on the KSC dataset. Notably, the model exhibited substantial improvements in challenging regions where spatial irregularities previously degraded classification performance. For instance, in the anisotropic "Stone–Steel–Towers" class, DTM improved classification accuracy by **56.7%** compared to the Mamba baseline (MiM). Similarly, for spectral-sensitive classes like "Water" and "Trees," DTM achieved a **7.2%** accuracy gain. These results highlight the importance of homogenizing spatial dependencies through frequency-domain regularization, which enhances the model's robustness to spatial variability and spectral similarity.

C. Complexity Analysis

While DTM's design incorporates multiple components to achieve high classification accuracy, it maintains competitive computational efficiency. The model required only **0.10G FLOPS** and **0.13M parameters**, significantly outperforming baseline models such as HiT, which uses 20.94M parameters and 11.93 FLOPS. Despite the added computational steps introduced by the DCT and HNN modules, DTM achieved a **50%** reduction in training time compared to the latest Mamba-based method (MiM), demonstrating that frequency-domain processing enhances spatial modeling efficiency without excessive computational overhead.

Table VIII reveals an important distinction between model size and computational latency. Although DTM achieves a 99% parameter reduction compared to Transformer baselines like HiT, its inference time is longer than that of a classic 2D-CNN. This phenomenon arises from the fundamental difference in architectural operations. The 2D-CNN leverages

TABLE VIII
COMPLEXITY ANALYSIS ON IPS DATASET

Method	FLOPS	Parameters (M)	Training Time (s)	Testing Time (s)
2D-CNN	0.20	0.32	**200.21**	**1.30**
3D-CNN	0.50	0.50	300.45	1.90
HybridSN	5.3	4.32	933.94	3.19
ViT	0.68	13.2	707.79	2.82
HiT	11.93	20.94	2005.29	8.77
MorphF	0.17	0.24	781.36	3.33
SSFTT	0.24	0.93	676.08	2.34
MiM	0.50	0.18	5344.58	20.94
MambaHSI	**0.01**	0.43	361.75	7.70
Ours	0.10	**0.13**	2679.64	11.17

highly optimized, purely local, and parallelizable convolutions. DTM, in its pursuit of effective long-range dependency modeling, necessarily incorporates components with higher sequential complexity: the selective SSM in Mamba blocks involves recurrent-like computations that limit parallelization, and the self-attention mechanism, despite our efficient design, introduces quadratic interactions. The DCTConv modules also add fixed per-forward-pass transformation costs. Therefore, DTM trades the parameter inefficiency of large CNNs or Transformers for higher operational complexity per parameter, resulting in a lightweight model that accurately captures global contexts but requires more sequential computation than a shallow local operator. This tradeoff is favorable in applications where model footprint is constrained (e.g., onboard deployment), and the significant accuracy gains justify the increased compute time per sample.

D. Limitations and Future Work

Despite its strong performance, DTM still has certain limitations. The current variance analysis focuses solely on spatial dimensions, which may limit performance in applications requiring temporal consistency. Future work will extend the use of ACF analysis to spectral–temporal dimensions, particularly for multitemporal datasets where the decay constant τ can vary across acquisition dates. In addition, we plan to investigate dynamic frequency-band selection to further optimize computational efficiency, as well as explore hardware-aware architecture designs to facilitate real-time, onboard remote sensing applications. Beyond current benchmarks, future evaluations will also incorporate modern datasets such as WHU-Hi to validate the scalability of DTM under higher-resolution and more diverse acquisition conditions.

VII. CONCLUSION

This article presented the DTM architecture, a novel approach to hyperspectral image (HSI) classification that leverages frequency-domain-guided feature learning to address critical spatial and spectral modeling challenges. We proposed a strategy to transform irregular and spatially inconsistent inputs into a more ordered and structurally coherent representation, thereby facilitating long-range dependency modeling. Our key contribution stems from the analysis of spatial dependency inconsistencies using the ACF, which motivated the integration of DCT-augmented processing with HNNs to enhance spatial coherence and achieve a 13.57% increase in the effective decay constant τ. To overcome the limitations of 1-D sequence modeling, multidirectional scanning mechanisms were introduced to better capture 2-D spatial dependencies. The hybrid DTM architecture enabled spectral–spatial disentanglement while significantly reducing model complexity, decreasing parameter count by 99%, from 20.94M in HiT to 0.13M. Extensive experiments demonstrated that DTM consistently achieves superior performance, reaching a state-of-the-art accuracy of 98.76% on the KSC dataset and showing a 5.69% improvement over Mamba under low-sample conditions. In addition, DTM reduced training time by 50% and operated efficiently with only 0.1G FLOPs, making it well-suited for real-time and onboard remote sensing applications. Beyond performance gains, DTM hyperspectral image analysis by integrating physically motivated frequency-domain representations with theoretically supported mechanisms that jointly stabilize spatial and spectral dependency modeling. Future work will explore dynamic frequency-band selection and hardware-aware design to further extend DTM's applicability in real-time remote sensing scenarios.

REFERENCES

[1] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag. Replaces Newslett.*, vol. 7, no. 2, pp. 101–117, Jun. 2019.

[2] B. Qin et al., "Hyperspherical structural-aware distillation enhanced spatial–spectral bidirectional interaction network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5524714.

[3] S. Feng, H. Zhang, B. Xi, C. Zhao, Y. Li, and J. Chanussot, "Cross-domain few-shot learning based on decoupled knowledge distillation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5534414.

[4] B. Qin, S. Feng, C. Zhao, B. Xi, W. Li, and R. Tao, "FDGNet: Frequency disentanglement and data geometry for domain generalization in cross-scene hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 6, pp. 10297–10310, Jun. 2025.

[5] Z. Wang, M. K. Ng, L. Zhuang, L. Gao, and B. Zhang, "Nonlocal self-similarity-based hyperspectral remote sensing image denoising with 3-D convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531617.

[6] M. Li, Y. Liu, G. Xue, Y. Huang, and G. Yang, "Exploring the relationship between center and neighborhoods: Central vector oriented self-similarity network for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1979–1993, Apr. 2023.

[7] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[8] R. A. Borsoi et al., "Spectral variability in hyperspectral data unmixing: A comprehensive review," *IEEE Geosci. Remote Sens. Mag. Replaces Newslett.*, vol. 9, no. 4, pp. 223–270, Dec. 2021.

[9] J. Theiler, A. Ziemann, S. Matteoli, and M. Diani, "Spectral variability of remotely sensed target materials: Causes, models, and strategies for mitigation and robust exploitation," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 8–30, Jun. 2019.

[10] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[11] J. X. Yang, J. Zhou, J. Wang, H. Tian, and A. W. C. Liew, "HSIMamba: Hyperpsectral imaging efficient feature learning with bidirectional state space for classification," 2024, *arXiv:2404.00272*.

[12] J. Hao, Y. Zhu, L. He, M. Liu, J. K. H. Tsoi, and K. F. Hung, "T-Mamba: A unified framework with long-range dependency in dual-domain for 2D & 3D tooth segmentation," 2024, *arXiv:2404.01065*.

[13] H. Pan, E. Hamdan, X. Zhu, A. E. Cetin, and U. Bagci, "Discrete cosine transform based decorrelated attention for vision transformers," 2024, *arXiv:2405.13901*.

[14] Y. He, B. Tu, B. Liu, J. Li, and A. Plaza, "3DSS-Mamba: 3D-spectral–spatial Mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5534216.

[15] X. Wu, Z. Liu, and L. Wang, "Spatio-temporal degradation model with graph neural network and structured state space model for remaining useful life prediction," *Rel. Eng. Syst. Saf.*, vol. 256, Apr. 2025, Art. no. 110770.

[16] K. Li et al., "VideoMamba: State space model for efficient video understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 237–255.

[17] W. Zhou, S.-I. Kamata, H. Wang, M.-S. Wong, Huiying, and Hou, "Mamba-in-Mamba: Centralized Mamba-cross-scan in tokenized Mamba model for hyperspectral image classification," 2024, *arXiv:2405.12003*.

[18] Q. Liu, J. Yue, Y. Fang, S. Xia, and L. Fang, "HyperMamba: A spectral–spatial adaptive Mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5538817.

[19] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-100, no. 1, pp. 90–93, 1974.

[20] M. Uličný, V. A. Krylov, and R. Dahyot, "Harmonic networks for image classification," in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 202. [Online]. Available: https://api.semanticscholar.org/CorpusID:203595213

[21] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

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

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

[24] Y. Li, Y. Luo, L. Zhang, Z. Wang, and B. Du, "MambaHSI: Spatial–spectral Mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5524216.

[25] J. Yao, D. Hong, C. Li, and J. Chanussot, "SpectralMamba: Efficient mamba for hyperspectral image classification," 2024, *arXiv:2404.08489*.

[26] J. Sheng, J. Zhou, J. Wang, P. Ye, and J. Fan, "DualMamba: A lightweight spectral–spatial Mamba-convolution network for hyperspectral image classification," 2024, *arXiv:2406.07050*.

[27] H. Hassani, N. Leonenko, and K. Patterson, "The sample autocorrelation function and the detection of long-memory processes," *Phys. A, Stat. Mech. Appl.*, vol. 391, no. 24, pp. 6367–6379, Dec. 2012.

[28] O. D. Faugeras and W. K. Pratt, "Decorrelation methods of texture feature extraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 4, pp. 323–332, Jul. 1980.

[29] C. Small, "High spatial resolution spectral mixture analysis of urban reflectance," *Remote Sens. Environ.*, vol. 88, nos. 1–2, pp. 170–186, Nov. 2003.

[30] S. Günther, L. Ruthotto, J. B. Schroder, E. C. Cyr, and N. R. Gauger, "Layer-parallel training of deep residual neural networks," 2018, *arXiv:1812.04352*.

[31] D. Uchaev and D. Uchaev, "Small sample hyperspectral image classification based on the random patches network and recursive filtering," *Sensors*, vol. 23, no. 5, p. 2499, Feb. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/5/2499

[32] Y. Zhang, W. Li, M. Zhang, S. Wang, R. Tao, and Q. Du, "Graph information aggregation cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1–14, Feb. 2024.

[33] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[34] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[35] X. Yang et al., "Synergistic 2D/3D convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 12, p. 2033, 2020.

[36] S. K. Roy et al., "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[37] 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.

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

[39] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.