remote sensing

MDPI

Article

MGWT: Hyperspectral Image Classification Based on a Multi-Granularity Weight Transformer Network

Zhuolin Xu [1], Wei Xiao [2], Anni Zhang [2], Siti Khairunniza Bejo [3], Ting Hu [4,5,6], Bin Zhou [7] and Shaoguang Huang [2,5,6,*]

1 School of Land Resources and Environment, Jiangxi Agricultural University, Nanchang 330045, China
2 School of Computer Science and Artificial Intelligence, China University of Geosciences, Wuhan 430074, China
3 Institute of Plantation Studies, Universiti Putra Malaysia, Serdang 43400, Malaysia
4 School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
5 Engineering Research Center of Natural Resource Information Management and Digital Twin Engineering Software, Ministry of Education, Wuhan 430074, China
6 Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430074, China
7 No. 722 Research Institute of China State Shipbuilding Corporation Limited, Wuhan 430205, China
* Correspondence: huangshaoguang@cug.edu.cn

Highlights

What are the main findings?

- The MGWT combines a token-level pathway with multi-scale group-proxy pathways for hyperspectral image classification.
- Spatial–spectral dual-stream attention and the LMTA jointly improve classification performance on three benchmark datasets.

What are the implications of the main findings?

- Multi-granularity aggregation captures fine-grained and progressively wider spatial–spectral supports within each attention block.
- The MGWT offers a robust approach for land-cover classification from hyperspectral remote sensing imagery.

Abstract

Transformer-based methods have achieved remarkable performance in hyperspectral image (HSI) classification. However, their ability to learn discriminative multi-scale spatial–spectral features remains limited by the single-granularity token-to-token interactions in conventional self-attention and the imbalanced modeling of spatial and spectral information. To address these limitations, we propose a multi-granularity weight transformer (MGWT) network comprising a spatial–spectral grouped parallel residual tokenizer (S^2-GPRT) and a multi-granularity attention (MGA) module. Specifically, the S^2-GPRT partitions the spectral bands into groups according to their physical characteristics and employs a dual-branch architecture to extract complementary local spatial–spectral features from individual spectral groups and the full spectrum. MGA integrates three components: a multi-granularity aggregator (Agg), a spatial–spectral dual-stream attention mechanism (S^2-DSA), and a learnable mixed top-k attention mechanism (LMTA). The Agg preserves a token-level pathway while constructing group-proxy pathways with progressively larger local receptive fields, enabling attention to integrate fine-grained information and multi-scale context within a single block. The S^2-DSA captures spatial and spectral dependencies through dedicated spatial and spectral attention streams. Furthermore, the LMTA constructs four rank-guided attention branches with different top-k selection ratios and adaptively fuses

Check for updates

Academic Editor: Wen Yang

Received: 4 September 2026
Revised: 21 October 2026
Accepted: 22 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

their outputs to integrate complementary attention responses at varying levels of sparsity. Experiments on three benchmark datasets demonstrate that the MGWT outperforms state-of-the-art methods in HSI classification.

Keywords: hyperspectral image; classification; CNN; transformer

1. Introduction

Hyperspectral imagery (HSI) combines spatial context with densely sampled spectra, helping to distinguish land-cover classes with similar appearances. These properties support applications in agricultural monitoring [1], urban planning [2], and target detection [3]. Hyperspectral image classification (HSIC) assigns a predefined class label to each pixel and supports these applications [4].

Early HSIC methods commonly separated feature extraction from classification. First, representative features were extracted from the HSI data. Second, these features were fed into classical machine learning classifiers, including *K*-nearest neighbors (KNN) [5], support vector machine (SVM) [6], random forest (RF) [7], and sparse representation classification (SRC) [8], for pixel-wise classification. Spectral-based methods exploit the raw spectral features or transform them into a lower-dimensional space by using techniques such as principal component analysis (PCA) [9], linear discriminant analysis (LDA) [10], or band selection [11,12]. Because they use only spectral features, these methods can be sensitive to noise and outliers. In contrast, spatial–spectral methods integrate spatial context by applying texture descriptors or filters (e.g., gray-level co-occurrence matrices [13] or a Gabor filter [14]) to extract spatial–spectral features, thereby significantly improving classification accuracy. However, these conventional methods rely on shallow feature extraction, which can limit generalization.

Deep learning methods have improved HSIC by learning nonlinear and hierarchical features from image data. Recurrent neural networks (RNNs) [15] model hyperspectral data as spectral sequences, where each band is treated as a time step to capture inter-band dependencies. Variants such as long short-term memory (LSTM) [16] and gated recurrent units (GRUs) [17] mitigate the gradient vanishing and exploding problems through gating mechanisms, thereby enhancing sequence modeling capability. In HSIC, GANs can generate high-quality virtual samples to alleviate the scarcity of labeled data and class imbalance, thus improving model generalization [18]. Graph convolutional networks (GCNs) [19] represent HSI as graph-structured data, where nodes correspond to pixel-wise spectral features and edges are constructed based on spatial proximity, spectral similarity, or hybrid strategies [20]. Through graph convolution operations, GCNs aggregate spatial–spectral information from local neighborhoods and capture long-range dependencies by stacking multiple layers. Convolutional neural networks (CNNs) are widely used because they capture local spatial patterns effectively. The representative models include 1D-CNN [21], 2D-CNN [22], and 3D-CNN [23], which focus on spectral, spatial, and joint spatial–spectral feature extraction, respectively. Hybrid architectures further improve performance by combining different network designs. For example, HybridSN [24] integrates 3D and 2D convolutions to effectively extract spatial–spectral features while reducing computational cost. Beyond the pure CNN designs, several studies combine CNNs with other architectures such as GCNs and RNNs to exploit complementary information. For instance, Liu et al. [25] integrated a multi-scale attention GCN with a dual-branch attention CNN to capture both long-range dependencies and local spatial–spectral features. Similarly, Li et al. [15] combined multi-scale CNNs with bidirectional RNNs to jointly model spatial and spectral information. In addition, advanced CNN architectures such as ResNet [26] and

DenseNet [27] introduce skip connections to alleviate gradient vanishing and information loss in deep networks. To further enhance feature discrimination, some methods incorporate auxiliary techniques, such as Gabor filtering [28] and Kullback–Leibler divergence-based constraints [29], to strengthen spatial or spectral feature learning. Moreover, attention mechanisms have been widely introduced into CNN-based frameworks [30–34]. For example, Zhu et al. [35] proposed a residual spatial–spectral attention network that adaptively reweights features across both channel and spatial dimensions, enabling more discriminative representation learning for HSIC.

Although CNNs capture local patterns effectively, small kernels require multiple layers to model long-range relationships. Vision transformers (ViTs) [36] use self-attention (SA) [37] to model long-range relationships across the spatial and spectral dimensions. However, global self-attention can overlook local feature details. A common strategy is to use CNN feature maps to improve tokenization or multi-head self-attention (MSA) [38–41]. For instance, the SSFTT [42] uses a hybrid CNN to extract features and perform tokenization. Fourierformer [43] designs a FourierMixer attention module that combines a CNN, MSA, and the Fourier transform to capture multi-scale features of spatial–spectral tokens. Standard ViTs form tokens from fixed-size image patches, which can limit multi-scale spatial representation. Jia et al. [44] proposed a six-branch heterogeneous fusion network with three spatial transformer branches and three spectral transformer branches, each extracting features on a different scale and optimizing the extraction of spatial–spectral features through dynamic weighted fusion. Fang et al. [45] constructed a dual-branch multi-scale transformer that employs deformable convolutions to generate different receptive fields for both spatial and spectral branches and incorporates a cross-branch pixel-level fusion mechanism. Since ViT's position encoding is added through random embedding and thus ignores the physical positional relationship of HSI pixels, Jiang et al. [46] proposed a graph generative structure-aware transformer that obtains positional priors of pixels by computing patch-label frequencies within sliding windows when generating position encoding. To mitigate information decay in deep ViT encoders, Sun et al. [47] designed MASSFormer, which uses a similar a priori strategy and introduces a memory tokenizer to convert spatial–spectral features into memory tokens that persistently store local context. To reduce ViT's parameter redundancy, Zhao et al. [48] proposed GSC-ViT, which decomposes standard MSA into grouped SA and point-wise SA, thereby lowering computational cost and reducing the number of parameters. Some CNN–transformer hybrids fuse branches by addition, concatenation, or simple attention, limiting direct feature exchange. The methods in [49,50] repeatedly pass global transformer context to CNN feature maps and local CNN features to token embeddings, enabling global-first or local-first representations. Peng et al. [51] further introduced a cross-SA mechanism to exchange spatial and spectral information between separate spatial-ViT and spectral-ViT branches.

Despite these advances, several challenges remain. First, the adopted SA operates within a single-scale feature space, where each head in MSA establishes pixel-level token-to-token correlations at a fixed granularity via the **Q**–**K**–**V** mechanism. This limits the ability of ViTs to effectively capture multi-scale spatial–spectral features of complex land cover. Second, most ViT-based HSIC methods primarily emphasize global spatial relationships among tokens while insufficiently modeling inter-channel (spectral) dependencies, which limits the discriminability of spatial–spectral features for classification. In addition, standard self-attention derives a single distribution from the complete score matrix, which may underrepresent complementary dependencies across different interaction ranges.

To address these limitations, we propose a multi-granularity weight transformer (MGWT) for HSIC, which mainly consists of a spatial–spectral grouped parallel residual tokenizer (S^2-GPRT) module, a multi-granularity attention (MGA) module, and a

Mixformer encoder. The S^2-GPRT module leverages the physical grouping prior of hyperspectral imaging to split spectral bands and adopts two parallel CNN branches to extract local spatial–spectral features within sub-bands and the entire band, respectively. The MGA module has three components: an aggregator, spatial–spectral dual-stream attention (S^2-DSA), and the LMTA. Unlike related multi-scale or grouped transformers that obtain scale diversity through parallel branches, hierarchical receptive fields, or groupwise-separable attention [43,45,48], the proposed aggregator operates inside each MGA block. One channel segment of each of **Q**, **K**, and **V** is retained, while three segments are aggregated with kernels of sizes 3, 5, and 7 before spatial and spectral attention. To compensate for the lack of inter-band (spectral) dependency modeling, the S^2-DSA explicitly distinguishes between spatial and spectral dimensions: the spatial stream captures spatial positional relations via standard SA, while the spectral stream uses a tailored SA variant to calculate spectral correlations. The LMTA constructs four attention branches from different score-ranking ranges in both streams and learns their relative contributions during training, thereby avoiding reliance on a manually selected single k. Finally, we replace the standard ViT encoder with a hierarchical Mixformer encoder to improve both classification accuracy and robustness. Our experimental results on three benchmark datasets demonstrate that the MGWT outperforms the existing state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 describes the materials and methods. Section 3 presents the experimental results. Section 4 discusses the main findings and practical limitations. Section 5 concludes the paper.

2. Materials and Methods

Figure 1 illustrates the overall framework of the proposed MGWT, which consists of three key modules: the S^2-GPRT module, the MGA module, and the Mixformer encoder. The S^2-GPRT module extracts shallow spatial–spectral information to generate discriminative tokens. The MGA module comprises three components. First, an aggregator combines an unchanged token-level pathway with group-proxy pathways formed at progressively larger local supports to capture multi-scale features. Second, the S^2-DSA component performs differentiated global modeling in the spatial and spectral dimensions based on the aggregator. Third, the LMTA constructs four attention branches with different top-k selection ranges and learns their contributions within the S^2-DSA. Finally, the Mixformer encoder adopts the proposed MGA and constructs a hierarchical feature extraction framework to enhance the representation learning.

2.1. Spatial–Spectral Grouped Parallel Residual Tokenizer

CNN–ViT hybrid methods for HSIC typically use convolutional layers to extract shallow features. For instance, the ACTN [50] utilizes 2D-CNN blocks to capture local spatial information of HSI, while [52] combines 3D-CNN blocks with transformation operations to generate spatial and spectral tokens, respectively. HSI data exhibits physical continuity between adjacent bands, indicating important local group structure in HSI. However, existing methods often indiscriminately process all the input bands, failing to incorporate such local spectral structure into the feature representation. Although recent studies [38,53] have adopted a grouped strategy by partitioning bands into sub-bands, they only extract features independently within each sub-band, thereby lacking holistic information across the entire spectral range.

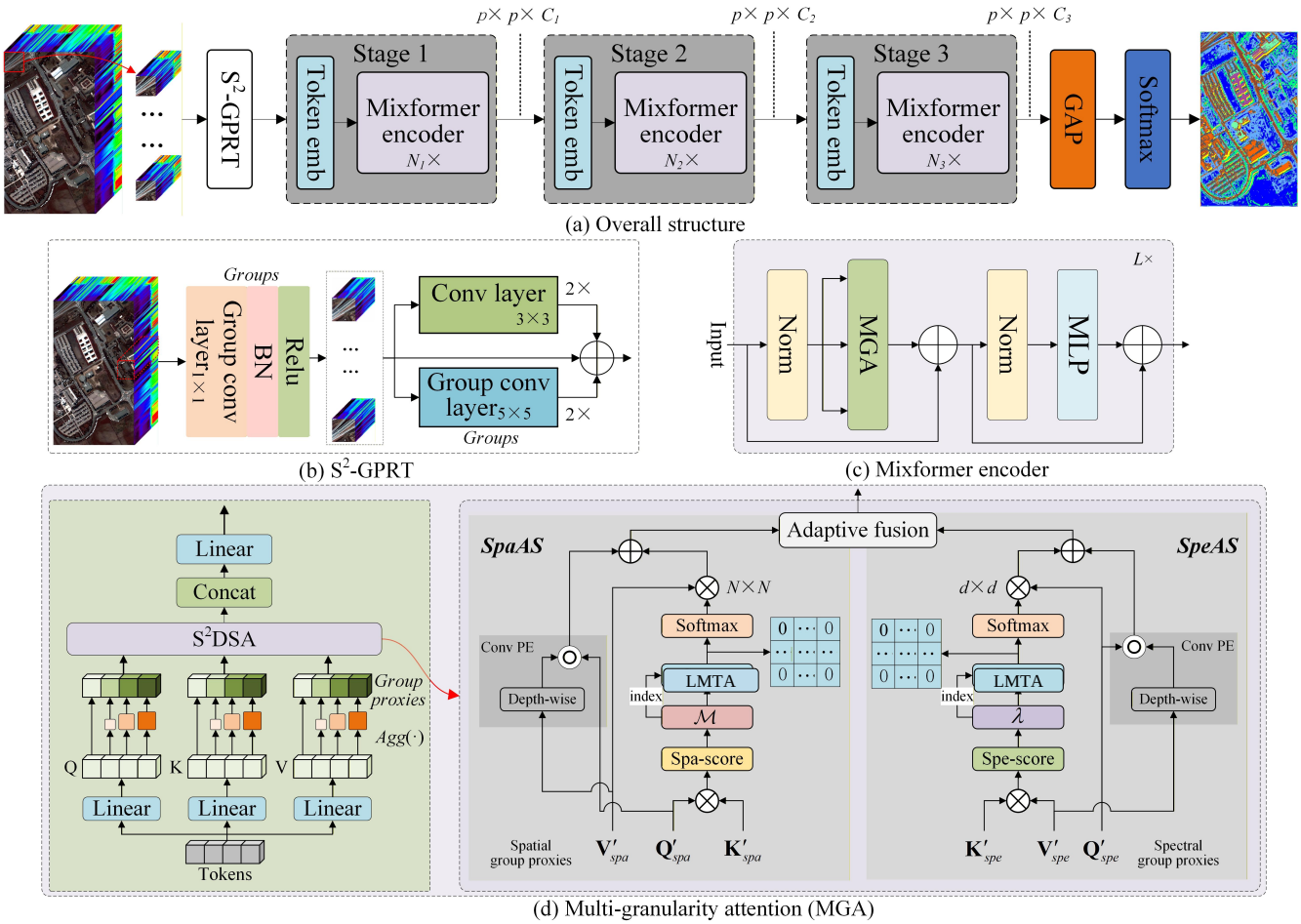


Figure 1. Overall framework of the proposed MGWT. (**a**) The model comprises a S^2-GPRT module, a three-stage hierarchical feature extraction network, and a classification head with global average pooling followed by a softmax layer. Each stage consists of a token embedding module and N_i Mixformer encoders. The token embedding module adjusts the output of the previous stage to match the token dimension of the current stage (i.e., C_1, C_2, and C_3). (**b**) The S^2-GPRT module extracts shallow multi-scale spatial–spectral features using grouped and standard convolutions. (**c**) The Mixformer encoder replaces standard MSA with the proposed MGA, thereby improving feature representation capability. (**d**) MGA first generates group proxies via an aggregator Agg(·), which are then fed into the S^2-DSA module. Within this module, the LMTA constructs and adaptively combines rank-guided attention branches in the spatial attention stream (SpaAS) and spectral attention stream (SpeAS).

To capture both types of information, we propose the S^2-GPRT module. As illustrated in Figure 1b, the S^2-GPRT first partitions all the spectral bands into multiple continuous sub-bands. Its two convolutional branches extract features from individual sub-bands and the full spectral range. This design preserves local band-group patterns while allowing information to pass across groups. Specifically, let $\mathbf{X} \in \mathbb{R}^{H\times W\times B}$ denote a hyperspectral image where H, W, and B are its height, width, and the number of spectral bands, respectively. Let $\mathbf{X}_i \in \mathbb{R}^{p\times p\times B}$ represent a local patch centered at the i-th pixel, with p denoting the size of the square window. We employ a 1×1 grouped convolution block $\mathrm{GConv}_{1\times1}$ consisting of a 1×1 grouped convolution layer, a batch normalization layer, and a ReLU activation to partition $\mathbf{X}_i$

into different band groups $\{\mathbf{X}_{i,1}, \mathbf{X}_{i,2}, \ldots, \mathbf{X}_{i,groups}\}$ and extract their corresponding spectral features. This local spectral feature extraction process can be formulated as follows:

$$\begin{aligned} \mathbf{X}'_i &= \mathrm{GConv}_{1\times1}(\mathbf{X}_i) \\ &= \{\mathbf{X}'_{i,1}, \mathbf{X}'_{i,2}, \ldots, \mathbf{X}'_{i,groups}\}, \end{aligned} \tag{1}$$

where $\mathbf{X}'_{i,j} \in \mathbb{R}^{p\times p\times b}$ is the j-th group of $\mathbf{X}'_i$, representing the extracted spectral features from the j-th band group $\mathbf{X}_{i,j}$, and $\mathbf{X'}_i$ is the output of the $\mathrm{GConv}_{1\times1}$ block. Here, b denotes the number of bands in each group after padding. When the spectral dimension is not divisible by the configured group number, the last band is replicated until equal grouping is possible; no band is discarded. Equation (1) only extracts the local spectral features among neighboring bands; in order to extract the global spectral features across all the bands, we design a dual-branch convolution block to perform multi-scale feature extraction on each $\mathbf{X}'_{i,j}$ and the overall spectral feature $\mathbf{X'}_i$.

A 5 × 5 two-dimensional convolution extracts spatial features from each $\mathbf{X}'_{i,j}$, and the resulting features are concatenated across band groups. This process is equivalent to the grouped convolution on $\mathbf{X}'_i$, i.e., $\mathrm{GConv}_{5\times5}(\mathbf{X}'_i)$. A second branch applies a 3 × 3 two-dimensional convolution to $\mathbf{X'}_i$, allowing information to pass across band groups. In both branches, batch normalization layers and ReLU activation functions are used. To alleviate information loss during the extraction of spatial–spectral features in both branches, a skip-connection branch is introduced, thereby achieving multi-scale feature aggregation. The feature fusion process of the three branches can be represented as follows:

$$\begin{aligned} \mathbf{T}_i &= \mathrm{ReLU}\big(\underbrace{\mathrm{Concat}\big(F(\mathbf{X}'_{i,1}), F(\mathbf{X}'_{i,2}), \ldots, F(\mathbf{X}'_{i,groups})\big)}_{\text{Sub-Band Features}} + \mathrm{Conv}_{3\times3}(\mathbf{X}'_i) + \mathbf{X}'_i\big) \\ &= \mathrm{ReLU}\big(\mathrm{GConv}_{5\times5}(\mathbf{X}'_i) + \mathrm{Conv}_{3\times3}(\mathbf{X}'_i) + \mathbf{X}'_i\big), \end{aligned} \tag{2}$$

where $F(\mathbf{X}'_{i,j})$ represents the output feature of the sub-band by using the 5 × 5 2D-convolutions, and $\mathrm{Conv}_{3\times3}(\mathbf{X'}_i)$ represents the feature obtained by processing all the bands with 3 × 3 2D-convolutions. After fusing the features from the three branches, the ReLU activation function is applied to introduce a nonlinear transformation. The S^2-GPRT therefore provides the Mixformer encoder with tokens that combine local band-group structure and full-spectrum spatial information.

2.2. Multi-Granularity-Based Self-Attention

To encode the multi-granularity spatial–spectral dependencies, we propose a novel self-attention model consisting of an aggregator, a S^2-DSA module, and the LMTA, which facilitates the discriminative capability of the learned features for HSIC.

2.2.1. Aggregator Agg(·)

Many ViT-based HSIC methods calculate attention between individual tokens. Such attention does not explicitly aggregate adjacent spatial tokens or contiguous spectral features into groups before computing their relationships, as illustrated in Figure 2a.

To address this problem, we propose an aggregator for HSI to model multi-granularity correlations in the ViT framework. Unlike conventional ViTs, which compute attention from unaggregated token representations, the proposed aggregator preserves one unaltered query–key–value segment and transforms the remaining segments into group proxies with different local supports. The subsequent attention calculation therefore combines fine-grained token evidence and multi-scale group-proxy evidence. We use this operational description instead of attributing a separately measured causal effect to each nominal interaction type. As illustrated in the left part of Figure 1d, the aggregator converts the

input $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ into group proxies. These proxies are obtained by uniformly splitting $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ along the channel dimension and applying distinct group aggregation operations. Let the input token sequence of MGA be denoted as $\mathbf{T}^i_{seq} \in \mathbb{R}^{N\times d}$, where $N = p \times p$ and d are the number of tokens and token dimension, respectively. $\mathbf{T}^i_{seq}$ is linearly projected through three learnable matrices $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v \in \mathbb{R}^{d\times d}$ to produce $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N\times d}$. The aggregator then splits $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ into S segments along the channel dimension, which can be expressed as

$$\mathbf{Q}^{(s)} = \mathbf{Q}[:, (s-1)\cdot\Delta_d : s\cdot\Delta_d], \quad \Delta_d = \left\lfloor \frac{d}{S} \right\rfloor, \tag{3}$$

where $\mathbf{Q}^{(s)} \in \mathbb{R}^{N\times\Delta_d}$ and $s = 1, 2, \ldots, S$. $\mathbf{K}^{(s)}$ and $\mathbf{V}^{(s)}$ are segmented in the same way. Following the split in Equation (3), the aggregator applies multi-scale aggregation to all but one segment, which is kept as an identity mapping to preserve the token-to-token scale correlation modeling. One-dimensional convolutions along the token or feature dimension form the spatial and spectral aggregators, respectively (Figure 3). Specifically, each aggregator employs three 1D convolutional kernels with sizes of 3, 5, and 7 to capture multi-scale features.

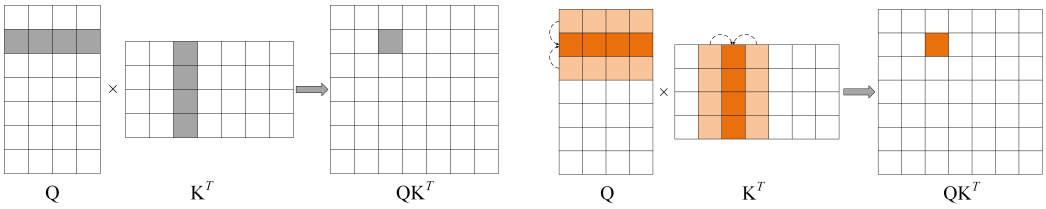


Figure 2. Granularity modeling within different self-attentions. (**a**) Token-to-token in the conventional methods. (**b**) Token groups to token groups in our method.

As shown in Figure 3a, the spatial aggregator employs a sliding window that moves along the token dimension, enabling aggregation of contextual information from adjacent token groups. The spatial aggregation operation for the s-th segment is then defined as

$$\hat{\mathbf{Q}}^{(s,k)}_{spa} = \mathrm{Conv1D}_k(\mathbf{Q}^{(s)}), \quad \text{stride} = 1, \quad \text{padding} = \left\lfloor \frac{k}{2} \right\rfloor, \tag{4}$$

where $\hat{\mathbf{Q}}^{(s,k)}_{spa} \in \mathbb{R}^{N\times\Delta_d}$ represents the output feature of the s-th token segment after passing through the spatial aggregator with 1D-convolution of kernel size k. The shared convolution kernel combines information from neighboring tokens. In the spectral aggregator, the sliding window slides along the feature dimension Δ_d to extract the cross-channel correlation within the same token so as to extract local spectral features. The spectral aggregation operation for the s-th segment is then defined as

$$\hat{\mathbf{Q}}^{(s,k)}_{spe} = \mathrm{Conv1D}_k(\mathbf{Q}^{(s)\top})^\top, \tag{5}$$

where $\hat{\mathbf{Q}}^{(s,k)}_{spe} \in \mathbb{R}^{N\times\Delta_d}$ represents the output features of the s-th token fragment processed by the k-th spectral aggregator.

After aggregation, the multi-scale aggregation results from each segmentation are concatenated along the feature dimension to generate the spatial group proxy and the spectral group proxy as follows:

$$\mathbf{Q}'_{spa} = \mathrm{Concat}\left(\hat{\mathbf{Q}}^{(1)}_{id1}, \hat{\mathbf{Q}}^{(2,3)}_{spa}, \hat{\mathbf{Q}}^{(3,5)}_{spa}, \hat{\mathbf{Q}}^{(4,7)}_{spa}\right), \tag{6}$$

where $\mathbf{Q}'_{spa} \in \mathbb{R}^{N\times d}$ represents the spatial group proxy, $\hat{\mathbf{Q}}^{(1)}_{id1}$ represents the identity mapping without spatial aggregation, and the number of segments S is set to 4.

$$\mathbf{Q}'_{spe} = \text{Concat}\left(\hat{\mathbf{Q}}^{(1)}_{id2}, \hat{\mathbf{Q}}^{(2,3)}_{spe}, \hat{\mathbf{Q}}^{(3,5)}_{spe}, \hat{\mathbf{Q}}^{(4,7)}_{spe}\right), \tag{7}$$

where $\mathbf{Q}'_{spe} \in \mathbb{R}^{N\times d}$ represents the spectral group proxy and $\hat{\mathbf{Q}}^{(1)}_{id2}$ represents the identity mapping without spectral aggregation. Similarly, $\mathbf{K}'_{spa}$, $\mathbf{V}'_{spa}$, $\mathbf{K}'_{spe}$, and $\mathbf{V}'_{spe}$ are generated following the same procedure.

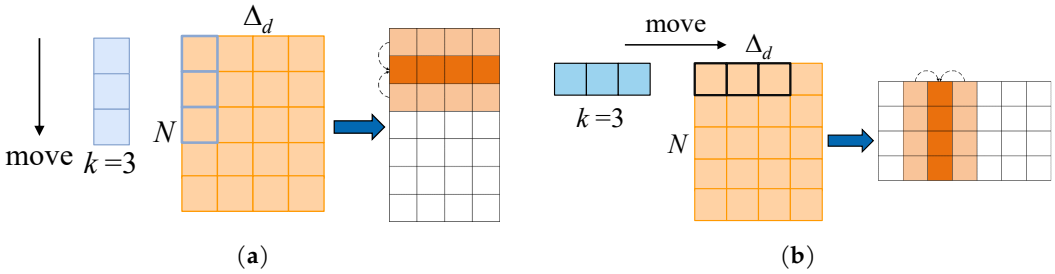


Figure 3. (**a**) Spatial aggregator. (**b**) Spectral aggregator.

The distinction from the existing grouped or multi-scale designs lies in where and how grouping is introduced. The aggregation is applied to the query, key, and value subspaces within every MGA block rather than only during input tokenization, in a separate multi-branch backbone, or solely to reduce attention cost. The identity segment preserves the original token resolution, whereas the 3-, 5-, and 7-element kernels create progressively wider spatial or spectral supports. The same set of multi-scale proxies is then processed by parallel $N \times N$ spatial attention and $d \times d$ spectral attention, which couples multi-scale aggregation with explicit spatial–spectral disentanglement.

2.2.2. The Spatial–Spectral Dual-Stream Attention (S^2-DSA) Module

Some ViT-based HSIC networks emphasize spatial dependencies while modeling spectral correlations less explicitly. Insufficient spectral modeling can make similar land-cover classes harder to distinguish. To address the problem, we propose the S^2-DSA module, which explicitly decouples the spatial and spectral dimensions in global SA modeling by capturing spatial and spectral dependencies separately, as shown on the right in Figure 1d. The S^2-DSA comprises two parallel streams: the SpaAS and the SpeAS. The SpaAS models spatial dependencies, while the SpeAS captures cross-band relationships.

Specifically, the SpaAS is formulated as follows:

$$\mathbf{A}_{\text{spa}} = \text{softmax}\left(\frac{\mathbf{Q}'_{\text{spa}}(\mathbf{K}'_{\text{spa}})^T}{\sqrt{d}}\right) \in \mathbb{R}^{N\times N}, \tag{8}$$

$$\mathbf{O}_{\text{spa}} = \mathbf{A}_{\text{spa}}\mathbf{V}'_{\text{spa}} \in \mathbb{R}^{N\times d}, \tag{9}$$

where $\mathbf{Q}'_{spa}$, $\mathbf{K}'_{spa}$, and $\mathbf{V}'_{spa}$ are the query, key, and value representations obtained from the spatial aggregator. Here, $\mathbf{A}_{spa}$ denotes the token-wise spatial dependency matrix, and $\mathbf{O}_{spa}$ represents the enhanced spatial features.

In parallel, the SpeAS can be formulated by

$$\mathbf{A}_{\text{spe}} = \text{softmax}\left((\mathbf{K}'_{\text{spe}})^T\mathbf{V}'_{\text{spe}}\right) \in \mathbb{R}^{d\times d}, \tag{10}$$

$$\mathbf{O}_{\text{spe}} = \frac{\mathbf{Q}'_{\text{spe}}}{\sqrt{d}}\mathbf{A}_{\text{spe}} \in \mathbb{R}^{N\times d}, \tag{11}$$

where $\mathbf{Q}'_{spe}$, $\mathbf{K}'_{spe}$, and $\mathbf{V}'_{spe}$ denote the query, key, and value derived from the spectral aggregator. The $\mathbf{A}_{spe}$ encodes global dependencies across bands, generating attention weights along the spectral dimension. This allows the model to enhance spectral feature discrimination, yielding enriched spectral representations $\mathbf{O}_{spe}$.

Building upon the SpaAS and SpeAS modules in Equations (8) and (10), we further incorporate a convolutional positional embedding (PE) into both streams to provide content-aware positional information. Unlike fixed absolute positional embeddings, this learnable depthwise convolution incorporates local spatial and spectral context.

The PE is fused with the original attention outputs through element-wise modulation with the query embeddings. Concretely, the spatial attention output $\mathbf{O}_{\text{spa}}$ in Equation (9) is reformulated as follows:

$$\mathbf{O}_{spa} = \mathbf{A}_{spa}\mathbf{V}'_{spa} + \mathbf{Q}'_{spa} \odot \text{Depthwise}(\mathbf{P}_1, \mathbf{V}'_{spa}), \tag{12}$$

where $\mathbf{P}_1$ denotes the weights of the depthwise convolution, and $\odot$ represents element-wise multiplication. The operator Depthwise$(\cdot)$ performs channel-wise local convolution on $\mathbf{V}'_{\text{spa}}$, producing a positional modulation map. This map is then combined with $\mathbf{Q}'_{\text{spa}}$ via element-wise multiplication to introduce position-aware corrections to the attention output. Similarly, the spectral attention output $\mathbf{O}_{\text{spe}}$ in Equation (11) is extended by incorporating a spectral positional term as follows:

$$\mathbf{O}_{spe} = \frac{\mathbf{Q}'_{spe}}{\sqrt{d}}\mathbf{A}_{spe} + \mathbf{Q}'_{spe} \odot \text{Depthwise}(\mathbf{P}_2, \mathbf{V}'_{spe}), \tag{13}$$

where $\mathbf{P}_2$ denotes the depthwise convolution kernel. The element-wise interaction between $\mathbf{Q}'_{spe}$ and the depthwise convolution output enforces a location-aware correction of spectral features.

By integrating the depthwise convolution PE into both SpaAS and SpeAS, the S^2-DSA module endows attention with learnable locality-preserving position cues, avoids global disturbances associated with random absolute encodings, and maintains the decoupled modeling of spatial and spectral dependencies, leading to improved local spatial–spectral feature extraction.

2.2.3. The Learnable Mixed Top-*k* Attention Mechanism

Standard self-attention generates one attention distribution from the complete score matrix. To provide complementary interaction ranges, the LMTA constructs four rank-guided branches in both SpaAS and SpeAS and learns their contributions during training. Let $\mathbf{S}$ denote a spatial or spectral score matrix and let L be the length of its last dimension. For branch m, the top-k count is determined by

$$k_m = \lfloor \rho_m L \rfloor, \qquad \rho_m \in \left\{ \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5} \right\}. \tag{14}$$

For the spatial and spectral streams, respectively, we define $\mathbf{P}_{\text{spa}} = \mathbf{Q}'_{\text{spa}}(\mathbf{K}'_{\text{spa}})^T/\sqrt{d}$ and $\mathbf{R}_{\text{spe}} = (\mathbf{K}'_{\text{spe}})^T\mathbf{V}'_{\text{spe}}$. The rank-guided score transformations for branch m are defined as

$$\begin{aligned} \mathcal{M}_{\rho_m}(\mathbf{P}_{\text{spa}})_{ij} &= \begin{cases} (\mathbf{P}_{\text{spa}})_{ij}, & j \in \text{TopKIndices}((\mathbf{P}_{\text{spa}})_{i,:}, k_m), \\ 0, & \text{otherwise}, \end{cases} \\ \lambda_{\rho_m}(\mathbf{R}_{\text{spe}})_{ij} &= \begin{cases} (\mathbf{R}_{\text{spe}})_{ij}, & j \in \text{TopKIndices}((\mathbf{R}_{\text{spe}})_{i,:}, k_m), \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{15}$$

where both operators retain the selected logits and map the remaining logits to a shared reference value of zero before normalization. The four attention branches are then calculated as

$$\mathbf{A}_{\text{spa}}^{(m)} = \text{softmax}\Big(\mathcal{M}_{\rho_m}(\mathbf{P}_{\text{spa}})\Big), \tag{16}$$

$$\mathbf{A}_{\text{spe}}^{(m)} = \text{softmax}\Big(\lambda_{\rho_m}(\mathbf{R}_{\text{spe}})\Big), \qquad m = 1, \ldots, 4. \tag{17}$$

Each branch uses a different selection range. Smaller ratios preserve fewer ranked logits, whereas larger ratios preserve more logits before the remaining entries are set to zero. Their outputs are combined using learnable coefficients as follows:

$$\mathbf{O}_{\text{spa},i} = \sum_{m=1}^{4} \alpha_m^{\text{spa}} \mathbf{A}_{\text{spa}}^{(m)} \mathbf{V}'_{\text{spa}} + \mathbf{Q}'_{\text{spa}} \odot \text{Depthwise}(\mathbf{P}_1, \mathbf{V}'_{\text{spa}}), \tag{18}$$

$$\mathbf{O}_{\text{spe},i} = \frac{1}{\sqrt{d}} \sum_{m=1}^{4} \alpha_m^{\text{spe}} \mathbf{Q}'_{\text{spe}} \mathbf{A}_{\text{spe}}^{(m)} + \mathbf{Q}'_{\text{spe}} \odot \text{Depthwise}(\mathbf{P}_2, \mathbf{V}'_{\text{spe}}). \tag{19}$$

Here, α_m^{spa} and α_m^{spe} are independent learnable parameters in each MGA block. Consequently, the four selection ratios are fixed and shared across datasets, while their contributions are learned from the training data. Because k_m is computed from L, the absolute selected counts automatically scale with the spatial token number or spectral feature dimension, and no single dataset-specific k is selected manually. We then extend the LMTA outputs into a multi-head representation as follows:

$$\mathbf{Out}_{\text{SpaAS}} = \text{Concat}\Big(\mathbf{O}_{\text{spa},1}, \mathbf{O}_{\text{spa},2}, \ldots, \mathbf{O}_{\text{spa},h}\Big), \tag{20}$$

$$\mathbf{Out}_{\text{SpeAS}} = \text{Concat}\Big(\mathbf{O}_{\text{spe},1}, \mathbf{O}_{\text{spe},2}, \ldots, \mathbf{O}_{\text{spe},h}\Big), \tag{21}$$

where h denotes the number of attention heads, and $\mathbf{O}_{spa,i}$ and $\mathbf{O}_{spe,i}$ are the features extracted by the i-th head in the SpaAS and SpeAS, respectively. $\mathbf{Out}_{SpaAS}$ and $\mathbf{Out}_{SpeAS}$ thus represent the aggregated attention outputs of all heads. To balance the contributions of the two streams, we propose an adaptive fusion mechanism that assigns learnable weights to spatial and spectral outputs,

$$\mathbf{Out}_{MGA} = \Big(w_{spa}\mathbf{Out}_{SpaAS} + w_{spe}\mathbf{Out}_{SpeAS}\Big)\mathbf{W}^T, \tag{22}$$

where w_{spa} and w_{spe} are learnable parameters to adaptively determine the relative importance of spatial and spectral features, $\mathbf{W} \in \mathbb{R}^{D\times D}$ with $D = h \times d$ is the standard MSA configuration, and $\mathbf{Out}_{MGA}$ serves as the fused spatial–spectral representation of the MGA block.

2.3. Mixformer Encoder and Hierarchical Structure

The Mixformer encoder incorporates MGA, as shown in Figure 1c. This encoder is employed to construct a three-stage hierarchical feature extraction network, as shown in Figure 1a. Each stage extracts spatial–spectral features at a different network depth.

Given the output of the S^2-GPRT module, $\mathbf{T}_i \in \mathbb{R}^{p\times p\times d}$, as shown in Equation (2), the j-th stage $j \in \{1, 2, 3\}$ contains a token emb module followed by N_j Mixformer encoders. The token emb applies a 1×1 convolution to adjust feature dimensions and flattens the result into tokens $\mathbf{T}_{seq,i}^{j} \in \mathbb{R}^{p^2\times C_j}$, with token dimensions changing across stages as $C_1 \to C_2 \to C_3$ to build progressively richer representations. The $\mathbf{T}_{seq,i}^{j}$ is processed by the Mixformer encoder equipped with MGA. Let $\mathbf{T}_{seq,i}^{j,0}$ denote the initial input to j-th-stage Mixformer encoder; this stage is formulated as follows:

$$\mathbf{Y}_{\text{seq},i}^{j,k} = \text{MGA}\Big(\text{LN}(\mathbf{T}_{\text{seq},i}^{j,k-1})\Big) + \mathbf{T}_{\text{seq},i}^{j,k-1}, \tag{23}$$

$$\mathbf{T}_{\mathrm{seq},i}^{j,k} = \mathrm{MLP}\Big(\mathrm{LN}(\mathbf{Y}_{\mathrm{seq},i}^{j,k})\Big) + \mathbf{Y}_{\mathrm{seq},i}^{j,k}, \tag{24}$$

where $k = 1, 2, \ldots, N_j$ and N_j denote the number of cascading Mixformer encoders within the j-th stage. $\mathbf{T}_{seq,i}^{j,k-1}$ denotes the output of the $(k-1)$-th Mixformer encoder and serves as the input to the k-th encoder. $\mathbf{Y}_{seq,i}^{j,k}$ represents the intermediate features processed by MGA, while $\mathbf{T}_{seq,i}^{j,k}$ is the final output of the k-th encoder at the j-th stage.

2.4. Classification Unit

The output of the three-stage hierarchical network, $\mathbf{T}_{seq,i}^{3,N_3} \in \mathbb{R}^{p^2 \times C_3}$, is fed to a classification unit composed of a global average pooling (GAP) layer, a fully connected (FC) layer and a softmax$(\cdot)$ function. Global average pooling summarizes each channel across spatial positions, yielding a descriptor $\mathbf{z} = [z_1, z_2, \ldots, z_{C_3}] \in \mathbb{R}^{C_3}$ computed as

$$z_c = \frac{1}{p^2} \sum_{u=1}^{p} \sum_{v=1}^{p} \mathbf{T}_{seq,i,u,v}^{3,N_3,c}, \tag{25}$$

where $\mathbf{T}_{seq,i,u,v}^{3,N_3,c}$ denotes the feature value at spatial position (u, v) of the c-th channel. The resulting vector $\mathbf{z}$ is then projected into the class space via the FC layer followed by a softmax function to produce the final class probabilities. We summarize our method in Algorithm 1.

Algorithm 1 Pseudocode of MGWT

Input: HSI data $\mathbf{X} \in \mathbb{R}^{H \times W \times B}$, ground truth $\mathbf{Y} \in \mathbb{R}^{H \times W}$
Output: Classification accuracy and classification map
Generate training and testing data.
for *epochs* **do**
    Feed $\mathbf{X}_i$ and extract $\mathbf{T}_i \in \mathbb{R}^{p \times p \times d}$ via S^2-GPRT
    Obtain tokens $\mathbf{T}_{seq,i} \in \mathbb{R}^{p^2 \times C_1}$ via Token emb
    Obtain $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ via linear projection
    Obtain $\mathbf{Q}'_{spa}, \mathbf{K}'_{spa}, \mathbf{V}'_{spa}$ via the spatial aggregator
    Obtain $\mathbf{Q}'_{spe}, \mathbf{K}'_{spe}, \mathbf{V}'_{spe}$ via the spectral aggregator
    Obtain $\mathbf{Out}_{spaAS}$ via SpaAS on $\mathbf{Q}'_{spa}, \mathbf{K}'_{spa}, \mathbf{V}'_{spa}$
    Obtain $\mathbf{Out}_{speAS}$ via SpeAS on $\mathbf{Q}'_{spe}, \mathbf{K}'_{spe}, \mathbf{V}'_{spe}$
    Obtain $\mathbf{Out}_{MGA}$ by fusing $\mathbf{Out}_{spaAS}$ and $\mathbf{Out}_{speAS}$
    Apply $\mathbf{Out}_{MGA}$ to residual connection, LN and MLP
    Obtain output of the first encoder block $\mathbf{T}_{seq,i}^{1,1}$
    Repeat process 6–14 for the remaining $N_1 - 1$ blocks
    Obtain $\mathbf{T}_{seq,i}^{1,N_1} \in \mathbb{R}^{p^2 \times C_1}$
    Obtain $\mathbf{T}_{seq,i}^{2,N_2} \in \mathbb{R}^{p^2 \times C_2}$ through stage 2 block
    Obtain $\mathbf{T}_{seq,i}^{3,N_3} \in \mathbb{R}^{p^2 \times C_3}$ through stage 3 block
    Apply classification unit
end for
Output: Classification map

We evaluate the MGWT on three benchmark HSI datasets: *Pavia University (PaviaU)*, *University of Houston (Houston2013)*, and *Xuzhou*.

2.5. Hyperspectral Datasets

Table 1 shows the basic information of the three HSI datasets.

Table 1. Spatial dimensions, spectral bands, labeled pixels, classes, and spectral ranges of the three HSI datasets.

Dataset	PaviaU	Houston2013	Xuzhou
Spatial size	610 × 340	349 × 1905	500 × 260
Bands	103	144	436
Labeled pixels	42,776	15,029	68,877
Classes	9	15	9
Spectral range	430–860 nm	380–1050 nm	–
Sensor	ROSIS	CASI	HYSPEX
Spatial resolution	1.3 m	2.5 m	0.73 m

2.5.1. PaviaU

This dataset was acquired in 2002 by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over the urban area of the University of Pavia, Italy. It comprises 610 × 340 pixels with 103 spectral bands. Figure 4a,b show the false-color composite and reference map, respectively; the map contains nine land-cover classes. The training and testing sample distribution is shown in Table 2.

Table 2. The number of training and test pixels for each class in *PaviaU*.

Class ID	Class Name	Train	Test
1	Asphalt	20	6611
2	Meadows	20	18,629
3	Gravel	20	2079
4	Trees	20	3044
5	Painted metal sheets	20	1325
6	Bare soil	20	5009
7	Bitumen	20	1310
8	Self-blocking bricks	20	3662
9	Shadows	20	927
	Total	180	42,596

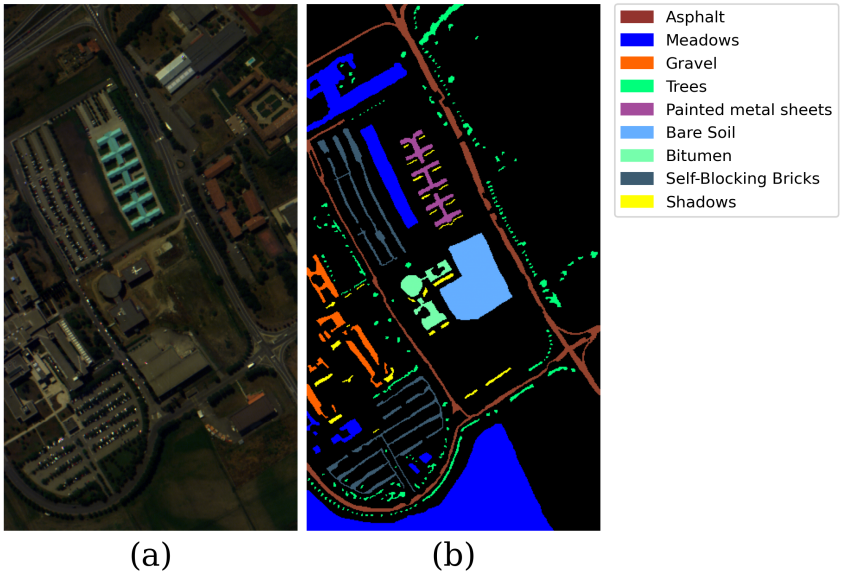


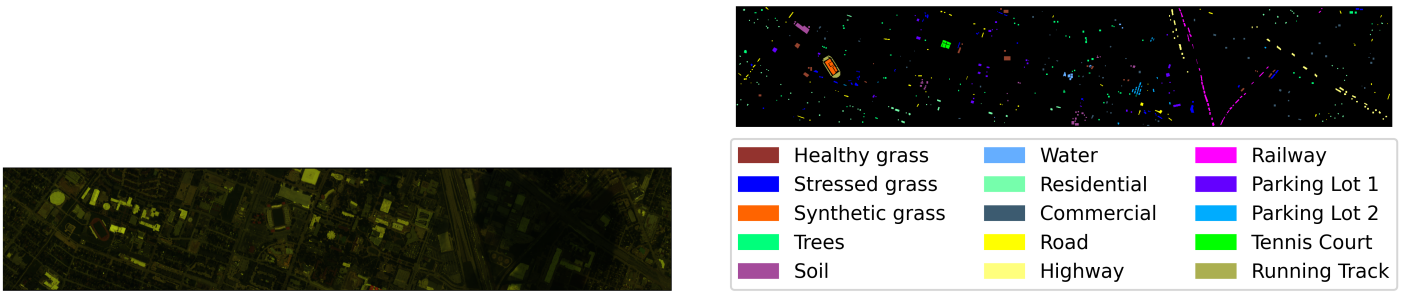
Figure 4. *PaviaU* scene. (**a**) False-color composite. (**b**) Reference land-cover labels.

2.5.2. Houston2013

This dataset was acquired by the ITRES-CASI 1500 sensor over the University of Houston campus and its surrounding urban areas. It was released as part of the 2013 IGRSS Data Fusion Contest. The dataset comprises 349 × 1905 pixels with 144 bands ranging from 380 nm to 1050 nm. Figure 5a,b show the false-color composite and reference map, respectively, and there are 15 land-cover classes in total. The distribution of training and testing samples for each class is summarized in Table 3.

Table 3. The number of training and test pixels for each class in *Houston2013*.

Class ID	Class Name	Train	Test
1	Grass healthy	20	1231
2	Grass stressed	20	1234
3	Grass synthetic	20	677
4	Tree	20	1224
5	Soil	20	1222
6	Water	20	305
7	Residential	20	1248
8	Commercial	20	1224
9	Road	20	1232
10	Highway	20	1207
11	Railway	20	1215
12	Parking lot 1	20	1213
13	Parking lot 2	20	449
14	Tennis court	20	408
15	Running track	20	640
	Total	300	14,729



(**a**) (**b**)

Figure 5. *Houston2013* scene. (**a**) False-color composite. (**b**) Reference land-cover labels.

2.5.3. Xuzhou

This dataset [54], released in 2019, was acquired by the airborne HYSPEX sensor over suburban areas of Xuzhou City, Jiangsu Province, China. The dataset comprises 500 × 260 pixels with a spatial resolution of 0.73 m and contains 436 bands. The scene is peri-urban and comprises 9 categories, including crops, vegetation, man-made structures, and coal fields. The false-color composite and reference map are shown in Figure 6a,b. The data split for the training set and testing set is shown in Table 4.

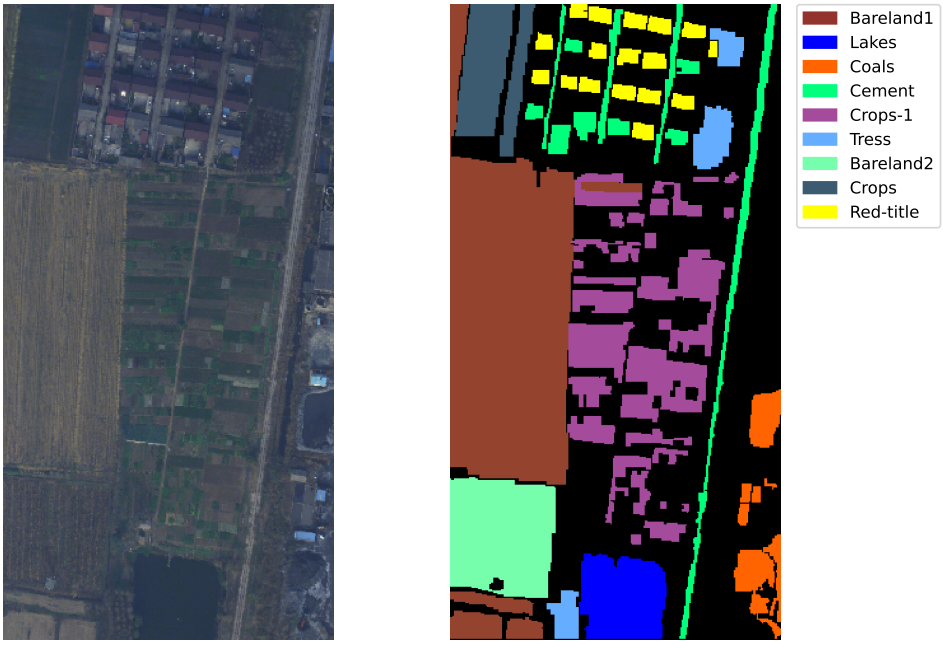


Figure 6. *Xuzhou* scene. (**a**) False-color composite. (**b**) Reference land-cover labels.

Table 4. The number of training and test pixels for each class in *Xuzhou*.

Class ID	Class Name	Train	Test
1	Bareland1	20	26,376
2	Lakes	20	4007
3	Coals	20	2763
4	Cement	20	5194
5	Crops-1	20	13,164
6	Trees	20	2416
7	Bareland2	20	6970
8	Crops	20	4757
9	Red-tile	20	3050
Total		180	68,697

2.6. Experiment Setting

2.6.1. Compared Methods

We compare the MGWT with two CNN-based methods, HybridSN [24] and DFFN [55], and seven transformer-based methods: SpectralFormer [56], HiT [57], SSFTT [42], CT-Mixer [58], GAHT [38], MorphFormer [39], and GSC-ViT [48]. HybridSN is a hybrid CNN model based on 3D-CNN and 2D-CNN. DFFN employs a deep CNN-based architecture that introduces residual learning to optimize deep network training and extract features at different depths, and GSC-ViT is a lightweight hyperspectral image classification method based on ViT.

2.6.2. Evaluation Metrics

We evaluate classification using per-class accuracy (ECA), overall accuracy (OA), average accuracy (AA), and Cohen's kappa coefficient (κ). OA refers to the ratio of the number of correctly classified pixels to the total number of pixels. AA is the average of the classification accuracy for each category, reflecting the ability of the classifier to recognize different categories, and κ is used to quantitatively assess the consistency of the predicted

results with the true label. Higher values indicate better performance for each metric. For a dataset with C category samples, let the number of samples in its c-th category be N_c ($1 \leq c \leq C$). The number of correctly categorized samples in that category is M_c ($M_c \leq N_c$). Every metric is calculated as follows:

$$OA = \frac{\sum_c M_c}{\sum_c N_c}, \tag{26}$$

$$AA = \frac{1}{C} \sum_c ECA_c, \tag{27}$$

$$\kappa = \frac{OA - P_e}{1 - P_e}, \tag{28}$$

where $ECA_c = \frac{M_c}{N_c}$ denotes the classification accuracy of the c-th class and $P_e = \frac{\sum_c M_c N_c}{(\sum_c N_c)^2}$.

2.6.3. Parameter Setting

The proposed model is implemented in Python 3.9.20. All the experiments run on a server equipped with an NVIDIA RTX 4090 GPU. For each dataset, 20 labeled pixels per class are randomly selected as training-sample centers and all the remaining labeled pixels form the test set; no separate validation split is used. Each input sample is a square patch centered on its corresponding labeled pixel, and the model settings are examined under this fixed train/test protocol as reported in Section 3.4. Detailed category information of the three datasets is shown in Tables 2–4. During training, the proposed MGWT is optimized using the AdamW optimizer with an initial learning rate of 0.001 and a weight decay of 0.05. A learning rate schedule combining a linear warm-up for the first 20 epochs and cosine annealing is adopted. The batch size is set to 64. All the experiments are conducted over ten independent runs, and the mean and standard deviation of each metric are reported.

3. Results

3.1. Performance Comparison

3.1.1. Quantitative Evaluation

We report the means and standard deviations of the classification accuracies of the different methods on the *PaviaU*, *Houston2013*, and *Xuzhou* datasets in Tables 5–7, respectively. The highest result in each row is bold, and the second highest is underlined. The MGWT achieves the highest OA, AA, and κ on PaviaU and Houston2013; on Xuzhou, its OA is close to that of CTMixer, while its AA is higher.

Among the CNN-based baselines, the DFFN achieves higher OA than HybridSN on all three datasets. Its deeper architecture may help to capture a broader spatial–spectral context, although these comparisons do not isolate the effect of network depth. Relative to the DFFN, the MGWT increases OA by 5.24, 5.81, and 2.99 percentage points on *PaviaU*, *Houston2013*, and *Xuzhou*, respectively. These comparisons are consistent with the value of combining local features and broader contextual information. It is worth noting, however, that not all the transformer-based methods outperform their CNN-based counterparts. For instance, HiT, SpectralFormer, and GSC-ViT generally yield lower accuracies than HybridSN and the DFFN. The main reasons are as follows: in HiT, the SA module is replaced with convolution operations, which restricts its ability to capture long-range spatial dependencies; SpectralFormer employs a simple tokenization strategy that neglects the inherent spatial structure of HSIs; and the lightweight design of GSC-ViT limits its capacity to model complex hyperspectral structures when training samples are scarce. By contrast, methods such as SSFTT, CTMixer, GAHT, and MorphFormer refine the tokenization and attention mechanisms, thereby achieving more competitive performance.

Table 5. Per-class accuracy and overall classification performance of the compared methods on *PaviaU*. The best result is marked in bold and the second-best result is underlined.

Class	HybridSN	DFFN	HiT	SpectralFormer	SSFTT	CTMixer	GAHT	Morph-Fomer	GSC-ViT	Proposed
1	83.77 ± 5.82	91.38 ± 0.42	52.04 ± 12.19	71.55 ± 6.62	89.34 ± 3.58	97.28 ± 0.69	92.36 ± 2.69	89.75 ± 3.03	87.20 ± 14.13	**97.53 ± 1.28**
2	93.53 ± 4.97	94.01 ± 2.44	69.48 ± 12.08	82.02 ± 4.34	94.42 ± 0.90	97.58 ± 0.58	92.51 ± 1.19	88.90 ± 6.83	74.98 ± 25.55	**98.88 ± 0.74**
3	94.62 ± 4.13	93.63 ± 2.55	78.32 ± 5.25	75.35 ± 7.33	94.96 ± 0.94	90.38 ± 2.38	80.78 ± 9.30	89.92 ± 7.46	85.13 ± 18.92	**98.67 ± 0.62**
4	83.37 ± 6.65	80.00 ± 2.52	93.55 ± 1.30	95.29 ± 1.72	95.96 ± 0.98	97.04 ± 0.36	**98.10 ± 0.64**	95.82 ± 0.80	95.34 ± 1.53	95.86 ± 0.45
5	99.43 ± 0.31	89.81 ± 4.67	99.52 ± 0.19	**100.0 ± 0.00**	99.70 ± 0.10	99.74 ± 0.08	99.98 ± 0.03	99.72 ± 0.17	99.49 ± 0.54	99.42 ± 0.08
6	94.36 ± 3.03	99.04 ± 0.59	79.47 ± 9.88	73.38 ± 7.68	99.09 ± 0.97	99.95 ± 0.06	90.84 ± 12.03	91.39 ± 5.18	86.09 ± 10.47	**99.99 ± 0.01**
7	99.79 ± 0.31	98.01 ± 1.01	78.19 ± 4.86	75.22 ± 5.20	99.98 ± 0.05	**99.99 ± 0.02**	93.40 ± 2.37	96.82 ± 1.00	74.44 ± 30.90	99.79 ± 0.18
8	66.73 ± 13.59	96.14 ± 0.69	72.63 ± 7.03	67.16 ± 10.18	78.12 ± 4.89	96.18 ± 1.35	89.63 ± 5.42	91.80 ± 4.22	57.89 ± 27.92	**97.96 ± 0.69**
9	83.52 ± 6.25	87.50 ± 4.39	87.81 ± 1.82	96.88 ± 1.41	97.04 ± 1.12	96.71 ± 0.30	**99.76 ± 0.20**	97.02 ± 1.04	98.68 ± 1.05	94.81 ± 0.93
OA	89.30 ± 3.13	93.21 ± 0.92	71.97 ± 4.99	79.40 ± 1.36	93.31 ± 1.08	97.42 ± 0.38	92.28 ± 1.92	90.87 ± 3.25	79.93 ± 14.25	**98.45 ± 0.38**
AA	88.79 ± 1.88	92.17 ± 0.58	79.00 ± 2.17	81.87 ± 1.29	94.29 ± 0.98	97.21 ± 0.41	93.04 ± 2.04	93.46 ± 1.52	84.36 ± 9.71	**98.10 ± 0.25**
κ	86.00 ± 3.92	91.08 ± 1.16	64.94 ± 5.42	73.40 ± 1.61	91.23 ± 1.40	96.60 ± 0.50	89.87 ± 2.57	88.17 ± 4.06	75.28 ± 16.21	**97.95 ± 0.50**

Table 6. Per-class accuracy and overall classification performance of the compared methods on *Houston2013*.

Class	HybridSN	DFFN	HiT	Spectral-Former	SSFTT	CTMixer	GAHT	Morph-Fomer	GSC-ViT	Proposed
1	89.76 ± 6.56	92.30 ± 1.38	82.49 ± 6.21	93.49 ± 2.60	94.96 ± 2.69	93.54 ± 2.73	94.32 ± 4.43	95.61 ± 1.73	83.00 ± 19.14	**97.51 ± 0.63**
2	91.77 ± 1.81	84.91 ± 3.08	71.10 ± 12.53	91.64 ± 3.75	93.30 ± 1.87	93.12 ± 0.64	90.46 ± 3.53	93.36 ± 4.07	79.06 ± 17.78	**95.62 ± 1.26**
3	98.54 ± 0.97	98.95 ± 0.33	91.23 ± 6.44	95.07 ± 1.81	99.38 ± 0.62	99.19 ± 0.12	99.01 ± 0.84	99.29 ± 0.46	96.65 ± 3.48	**99.60 ± 0.26**
4	78.26 ± 7.01	83.13 ± 3.20	70.56 ± 8.56	84.57 ± 6.48	**95.27 ± 1.43**	93.56 ± 0.66	88.33 ± 2.58	91.39 ± 1.84	82.30 ± 14.76	93.91 ± 0.94
5	94.53 ± 4.78	95.05 ± 5.29	92.13 ± 3.84	96.42 ± 1.87	96.08 ± 3.34	**99.56 ± 0.28**	96.48 ± 1.97	91.46 ± 8.38	83.45 ± 18.05	99.44 ± 0.43
6	99.48 ± 0.83	99.31 ± 0.77	85.08 ± 3.60	93.97 ± 3.88	**100.0 ± 0.00**	**100.0 ± 0.00**	96.52 ± 2.04	97.08 ± 1.16	92.56 ± 9.17	99.87 ± 0.22
7	77.30 ± 8.85	65.22 ± 3.81	66.95 ± 9.35	71.97 ± 5.08	77.50 ± 8.96	**86.33 ± 1.29**	76.83 ± 4.62	83.71 ± 2.15	59.10 ± 21.68	84.41 ± 2.89
8	63.31 ± 5.58	63.49 ± 4.01	38.80 ± 7.77	66.18 ± 3.41	63.42 ± 4.80	71.48 ± 1.98	66.79 ± 2.39	66.40 ± 4.51	**77.37 ± 7.62**	72.18 ± 1.77
9	63.29 ± 7.71	78.98 ± 3.37	68.78 ± 13.30	60.82 ± 10.95	79.84 ± 4.38	**88.24 ± 1.66**	79.26 ± 5.62	81.47 ± 7.13	56.29 ± 25.96	87.55 ± 1.23
10	86.64 ± 4.63	96.13 ± 3.96	65.78 ± 9.24	81.92 ± 4.68	92.44 ± 3.95	96.89 ± 1.48	88.85 ± 13.34	88.91 ± 11.38	51.77 ± 20.69	**99.55 ± 0.45**
11	90.82 ± 3.96	93.07 ± 3.29	48.82 ± 19.06	65.21 ± 6.48	**93.70 ± 3.21**	91.95 ± 3.12	77.29 ± 5.63	86.32 ± 3.08	43.42 ± 22.89	89.45 ± 1.18
12	92.41 ± 4.03	**92.69 ± 1.05**	41.57 ± 21.87	66.74 ± 8.14	92.69 ± 1.76	91.70 ± 1.60	85.42 ± 5.92	83.78 ± 6.55	69.23 ± 19.05	92.65 ± 1.60
13	87.28 ± 3.64	88.86 ± 1.74	74.61 ± 21.01	68.55 ± 11.01	84.92 ± 1.95	**90.62 ± 1.97**	86.84 ± 6.27	84.54 ± 4.99	61.96 ± 25.73	88.91 ± 2.68
14	**100.0 ± 0.00**	**100.0 ± 0.00**	97.43 ± 1.95	96.86 ± 2.32	**100.0 ± 0.00**	99.93 ± 0.22	98.92 ± 0.61	99.36 ± 1.10	84.88 ± 29.06	**100.0 ± 0.00**
15	98.84 ± 1.33	95.89 ± 1.62	91.56 ± 3.22	98.64 ± 1.34	99.88 ± 0.15	**100.0 ± 0.00**	99.50 ± 0.88	98.45 ± 2.21	83.13 ± 23.72	98.91 ± 0.70
OA	85.16 ± 1.51	86.47 ± 0.94	68.75 ± 2.20	80.17 ± 3.13	89.43 ± 1.72	91.87 ± 0.26	86.45 ± 2.07	87.90 ± 1.82	71.21 ± 8.01	**92.28 ± 0.37**
AA	87.48 ± 1.19	88.53 ± 0.82	72.46 ± 2.63	82.13 ± 3.04	90.89 ± 1.41	93.07 ± 0.22	88.32 ± 1.81	89.41 ± 1.44	73.61 ± 9.31	**93.30 ± 0.35**
κ	83.96 ± 1.62	85.38 ± 1.02	66.27 ± 2.40	78.57 ± 3.39	88.57 ± 1.86	91.21 ± 0.28	85.36 ± 2.22	86.93 ± 1.96	69.01 ± 8.51	**91.65 ± 0.39**

Table 7. Per-class accuracy and overall classification performance of the compared methods on *Xuzhou*.

Class	HybridSN	DFFN	HiT	Spectral-Former	SSFTT	CTMixer	GAHT	Morph-Fomer	GSC-ViT	Proposed
1	87.54 ± 2.24	93.26 ± 2.42	86.07 ± 2.51	87.99 ± 1.05	91.25 ± 1.25	**95.89 ± 0.59**	93.31 ± 1.58	93.16 ± 1.68	88.12 ± 5.81	94.85 ± 1.04
2	**99.96 ± 0.06**	99.62 ± 0.29	95.40 ± 1.23	94.27 ± 0.31	99.71 ± 0.14	99.59 ± 0.17	99.04 ± 0.26	99.51 ± 0.18	98.55 ± 1.04	99.92 ± 0.08
3	92.16 ± 7.19	97.14 ± 2.50	80.34 ± 3.71	72.95 ± 2.04	98.14 ± 0.63	99.08 ± 0.36	94.76 ± 1.03	95.00 ± 2.49	93.81 ± 5.20	**99.52 ± 0.45**
4	87.67 ± 3.48	86.09 ± 4.10	75.91 ± 6.42	67.49 ± 8.10	92.77 ± 2.06	94.61 ± 0.98	91.17 ± 0.72	90.78 ± 2.64	76.31 ± 24.13	**96.06 ± 0.76**
5	89.95 ± 4.94	89.98 ± 1.70	92.06 ± 3.52	82.00 ± 3.28	**97.71 ± 1.05**	96.82 ± 0.90	94.71 ± 0.70	92.77 ± 2.57	86.86 ± 13.12	97.30 ± 0.34
6	93.81 ± 3.77	98.16 ± 1.00	86.33 ± 5.31	78.92 ± 3.32	**99.05 ± 1.26**	96.85 ± 0.97	93.04 ± 1.66	93.20 ± 3.10	66.72 ± 33.00	96.57 ± 0.93
7	96.77 ± 2.36	98.82 ± 0.30	94.96 ± 6.00	87.76 ± 7.13	99.40 ± 0.29	99.06 ± 0.19	94.88 ± 1.34	98.78 ± 0.50	72.35 ± 33.28	**99.55 ± 0.09**
8	99.69 ± 0.54	**99.93 ± 0.09**	96.95 ± 1.70	92.22 ± 0.47	98.43 ± 0.65	98.67 ± 0.30	98.44 ± 0.22	98.07 ± 1.06	86.94 ± 17.67	99.16 ± 0.37
9	98.50 ± 1.20	96.67 ± 0.93	93.59 ± 1.35	77.61 ± 9.77	99.45 ± 0.39	99.04 ± 0.72	98.93 ± 0.48	97.51 ± 1.47	94.26 ± 6.73	**99.93 ± 0.03**
OA	91.41 ± 1.49	93.97 ± 1.08	88.76 ± 2.47	84.54 ± 1.41	95.34 ± 0.52	**97.00 ± 0.29**	94.56 ± 0.58	94.45 ± 0.37	85.66 ± 5.30	96.96 ± 0.44
AA	94.00 ± 1.54	95.52 ± 0.54	89.07 ± 2.53	82.36 ± 2.33	97.32 ± 0.42	97.73 ± 0.20	95.36 ± 0.28	95.42 ± 0.44	84.88 ± 9.20	**98.10 ± 0.26**
κ	89.27 ± 1.85	92.39 ± 1.35	85.97 ± 3.05	80.66 ± 1.82	94.14 ± 0.64	**96.21 ± 0.37**	93.14 ± 0.72	93.00 ± 0.45	81.81 ± 7.04	96.16 ± 0.55

Compared with the transformer-based methods, the MGWT achieves the highest OA, AA, and κ on PaviaU and Houston2013 and remains competitive on Xuzhou. On the *Pavia* dataset, the MGWT surpasses the second-best method, CTMixer, by 1.03%, 0.89%, and 1.35% in OA, AA, and κ, respectively. The same ranking is observed on *Houston2013*. On the *Xuzhou* dataset, the MGWT performs comparably to CTMixer while achieving a 0.47% improvement in AA, indicating that the MGWT not only possesses competitive overall classification ability but also demonstrates better class balance. As shown in Table 7, the MGWT attains the best or second-best ECA in seven out of nine categories. The class-level results further show where the aggregate improvement occurs. On *PaviaU*, the clearest gains over the strongest competing class result are obtained for gravel and self-blocking bricks, with increases from 94.96% to 98.67% and from 96.18% to 97.96%, respectively. On *Houston2013*, the MGWT improves grass healthy, grass stressed, and highway from 95.61%, 93.36%, and 96.89% to 97.51%, 95.62%, and 99.55%, respectively. On *Xuzhou*, the model is the strongest for coals, cement, Bareland2, and red-tile. Together with the smoother regions and cleaner boundaries in the classification maps, these results indicate that the principal reductions concern isolated-pixel noise, fragmented boundaries, and confusion among visually or spectrally similar land-cover categories. These observations support the value of combining fine-grained and grouped context at the model level, but they do not isolate a separate causal contribution for each interaction pathway.

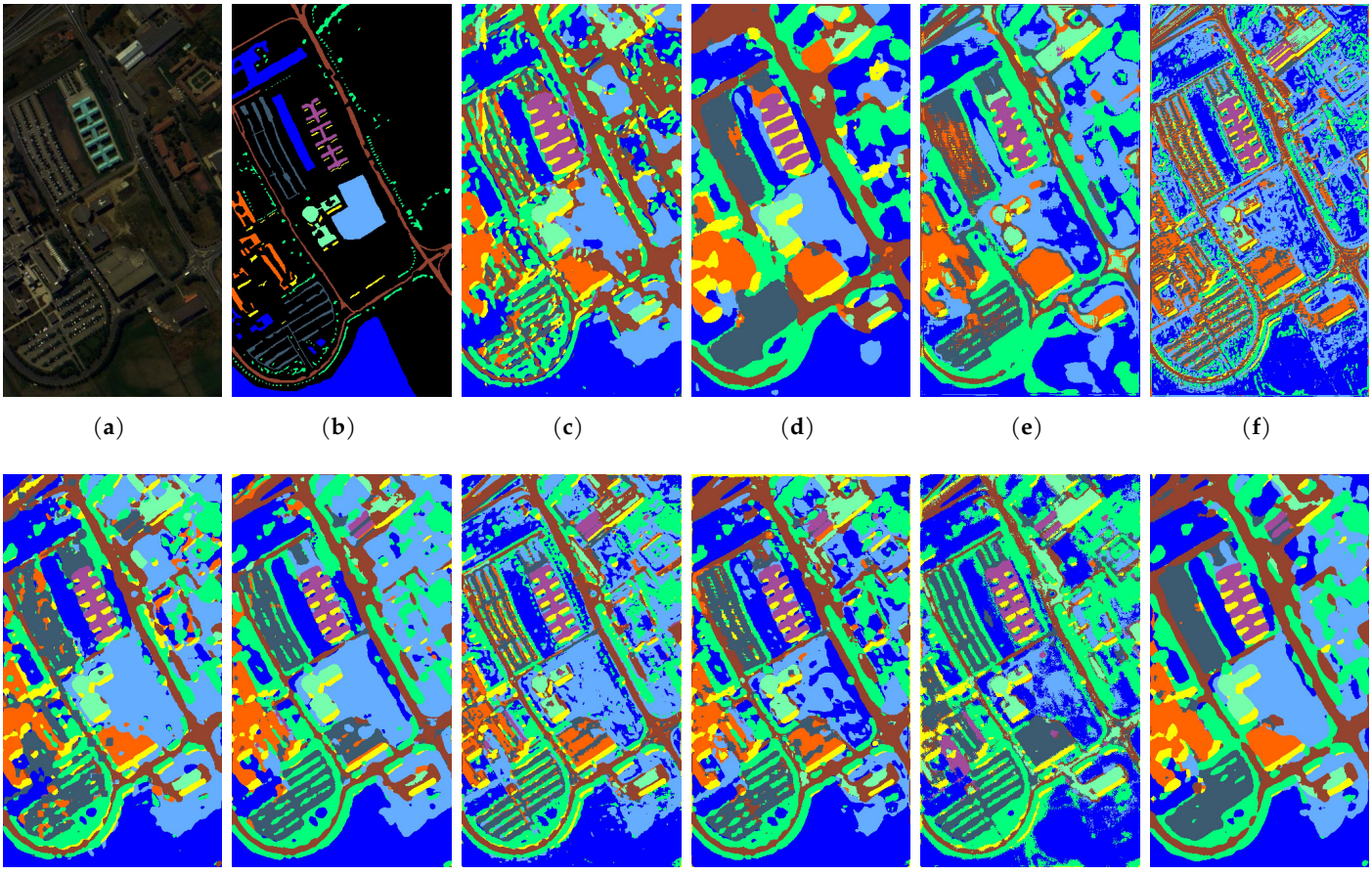
3.1.2. Visual Evaluation

Figures 7–9 compare the pixel-level classification maps produced by all the methods. The MGWT produces fewer isolated misclassified pixels across the three scenes. For the *Pavia* dataset, as shown in Figure 7, the classification map produced by the MGWT exhibits clearer and smoother boundaries in areas such as buildings and roads, and regions belonging to the same class are more coherent. On the *Houston2013* dataset, Figure 8 further demonstrates that the MGWT yields fewer misclassified regions. For the *Xuzhou* dataset, Figure 9 shows that the MGWT not only produces smooth classification results but also preserves fine details, particularly along the boundaries between different land-cover types. For example, although the MGWT does not achieve the highest OA, its classification map exhibits clearer boundaries in road regions labeled by green pixels (e.g., the long straight road on the right side of the image) compared with the method CTMixer, which attains the highest OA. Moreover, by comparing with the false-color image, the MGWT is the only method that can accurately delineate the contour of the unlabeled dark region on the right side of the road using blue pixels. This observation indicates that the MGWT is able to better exploit spatial information in hyperspectral classification tasks, achieving high classification accuracy while preserving fine-grained structural details. Overall, the maps provide visual context for the quantitative comparisons, particularly at class boundaries.

3.2. Ablation Study

The S^2-GPRT and MGA modules are key components of the proposed MGWT framework. To investigate their contributions to the overall performance, ablation studies are conducted on both modules. For the S^2-GPRT module, the 1×1 grouped convolution layer and the grouped convolution branch are removed while retaining only the standard 2D-CNN convolution branch. This setting is designed to evaluate the effectiveness of incorporating the physical grouping characteristics of hyperspectral images for enhancing spatial–spectral feature extraction. For the MGA module, the MGWT is evaluated under several simplified variants to analyze the impact of its components, including the aggregator, the S^2-DSA (i.e., the SpaAS and the SpeAS), and the LMTA. In the ablation study of the aggregator, all the segment-wise operations on $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are replaced with identity

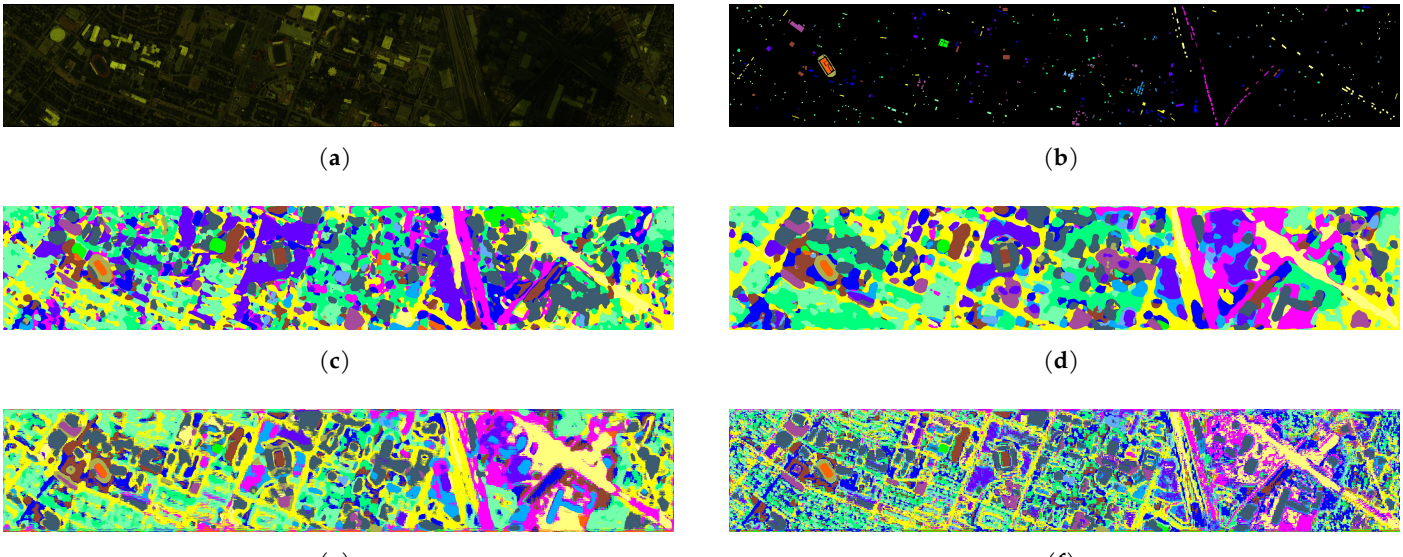
mappings, causing the group proxies to degenerate into the standard token representations used in conventional ViTs. The full configuration of the S^2-GPRT and MGA is taken as the baseline, and the ablation results are reported in Table 8.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

(**g**) (**h**) (**i**) (**j**) (**k**) (**l**)

Figure 7. *Pavia* image. (**a**) False color, (**b**) ground truth, (**c**) HybridSN (OA = 90.45%), (**d**) DFFN (OA = 93.67%), (**e**) HiT (OA = 71.14%), (**f**) SpectralFormer (OA = 78.57%), (**g**) SSFTT (OA = 94.54%), (**h**) CTMixer (OA = 97.43%), (**i**) GAHT (OA = 92.91%), (**j**) MorphFormer (OA = 92.94%), (**k**) GSC-ViT (OA = 80.36%), and (**l**) our MGWT (OA = 98.42%).

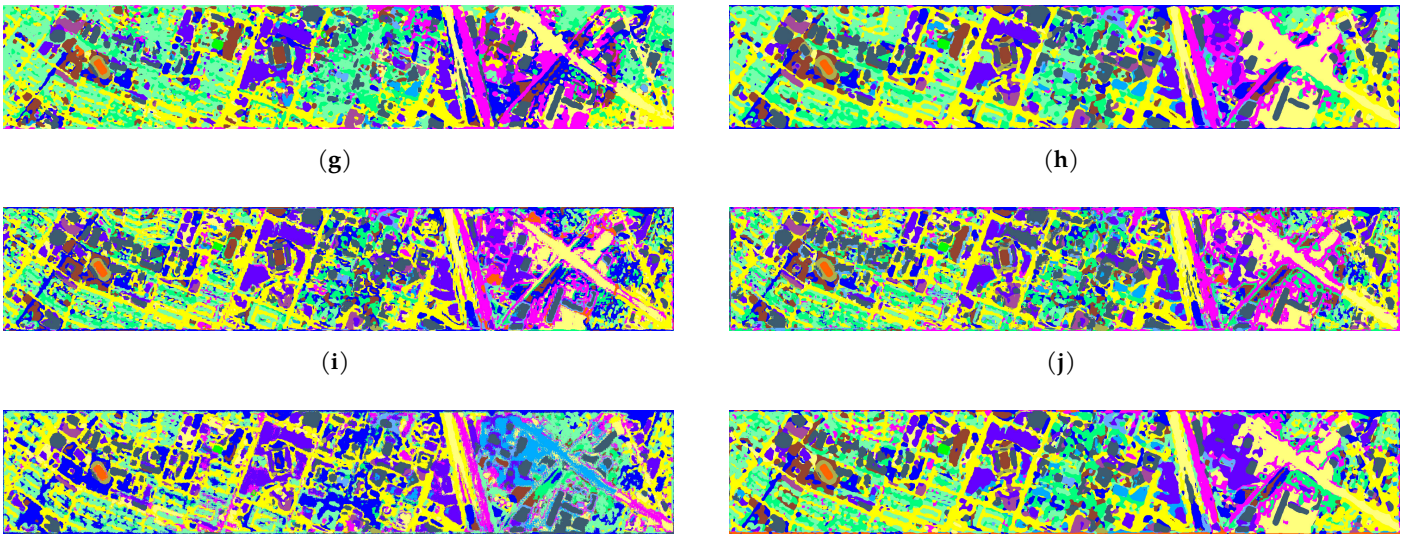


(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 8. *Cont*.

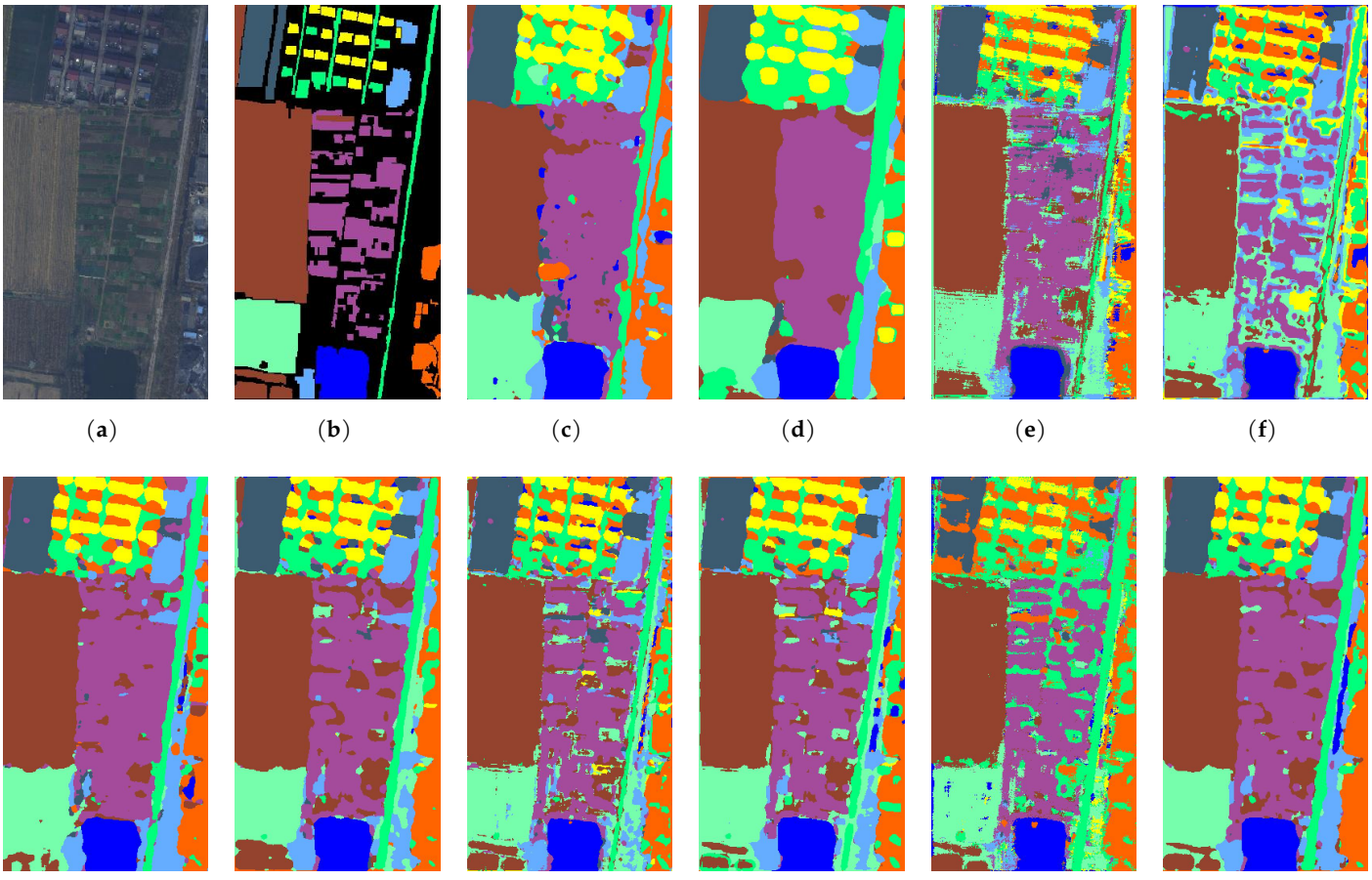


(**g**) (**h**)

(**i**) (**j**)

(**k**) (**l**)

Figure 8. *Houston2013* image. (**a**) False color, (**b**) ground truth, (**c**) HybridSN (OA = 86.45%), (**d**) DFFN (OA = 86.20%), (**e**) HiT (OA = 68.78%), (**f**) SpectralFormer (OA = 82.40%), (**g**) SSFTT (OA = 89.30%), (**h**) CTMixer (OA = 91.91%), (**i**) GAHT (OA = 86.37%), (**j**) MorphFormer (OA = 89.16%), (**k**) GSC-ViT (OA = 71.40%), and (**l**) our MGWT (OA = 92.30%).



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

(**g**) (**h**) (**i**) (**j**) (**k**) (**l**)

Figure 9. *Xuzhou* image. (**a**) False color, (**b**) ground truth, (**c**) HybridSN (OA = 91.39%), (**d**) DFFN (OA = 94.01%), (**e**) HiT (OA = 88.45%), (**f**) SpectralFormer (OA = 84.47%), (**g**) SSFTT (OA = 95.35%), (**h**) CTMixer (OA = 96.99%), (**i**) GAHT (OA = 94.49%), (**j**) MorphFormer (OA = 94.00%), (**k**) GSC-ViT (OA = 86.23%), and (**l**) our MGWT (OA = 96.96%).

Table 8. Ablation results for MGWT components on the three benchmark datasets.

Schemes	S^2-GPRT	MGA				OA		
		Agg	SpaAS	SpeAS	LMTA	PaviaU	Houston	Xuzhou
MGWT	✓	✓	✓	✓	✓	**98.45 ± 0.38**	**92.28 ± 0.37**	**96.96 ± 0.44**
1	2D-CNN	✓	✓	✓	✓	95.57 ± 0.95 (−2.88)	90.82 ± 0.23 (−1.47)	96.49 ± 0.52 (−0.47)
2	✓	✓	✓	✓	×	97.79 ± 0.91 (−0.66)	91.81 ± 0.41 (−0.47)	96.35 ± 0.70 (−0.61)
3	✓	✓	×	✓	✓	97.90 ± 0.38 (−0.55)	91.27 ± 0.43 (−1.01)	95.98 ± 0.61 (−0.98)
4	✓	✓	✓	×	✓	97.54 ± 0.57 (−0.91)	91.65 ± 0.40 (−0.63)	96.58 ± 0.63 (−0.38)
5	✓	×	✓	✓	✓	98.16 ± 0.33 (−0.29)	91.84 ± 0.33 (−0.44)	96.49 ± 0.56 (−0.47)

Relative to the complete MGWT model, Scheme 1 has lower OA on all three datasets, supporting a contribution from the S^2-GPRT module. Removing the LMTA in Scheme 2 reduces OA by 0.66, 0.47, and 0.61 percentage points on *PaviaU*, *Houston2013*, and *Xuzhou*, respectively. The moderate but consistent decreases indicate that the LMTA acts as a complementary attention-refinement component rather than the dominant source of the MGWT's performance. The S^2-GPRT, multi-granularity aggregator, spatial–spectral dual-stream attention, and hierarchical encoder retain the principal feature-representation capability. The LMTA provides an additional improvement by constructing attention branches over different score-selection ranges and learning their relative contributions. Schemes 3 and 4 validate the effectiveness of the S^2-DSA mechanisms, respectively, with the results suggesting that the spatial-stream attention contributes more substantially to overall classification performance. Scheme 5 shows varying degrees of OA reduction on all three datasets, further confirming the effectiveness of the aggregator in feature fusion and information extraction.

3.3. Computational Complexity Analysis

We further compare the computational complexity of the MGWT with the competing methods on all three datasets. Table 9 reports the number of floating-point operations (FLOPs) and model parameters. FLOPs are reported in millions (M), and the numbers of parameters are reported in thousands (K).

Table 9. Computational complexity of different methods on the three datasets.

Dataset	Metric	HybridSN	DFFN	HiT	SpectralFormer	SSFTT	CTMixer	GAHT	MorphFormer	GSC-ViT	MGWT
PaviaU	FLOPs (M)	7272.35	11,317.30	74,091.92	**1766.41**	2052.36	16,951.25	31,035.55	4118.40	1988.26	211,979.24
	Params (K)	2369.66	491.69	48,218.22	231.02	153.22	689.83	927.11	226.86	**152.71**	3386.55
Houston2013	FLOPs (M)	390.09	3798.96	17,184.38	2358.55	**240.43**	3445.33	5530.44	1156.01	807.55	17,795.88
	Params (K)	403.58	492.80	51,175.17	241.65	**153.22**	657.33	973.58	232.69	<u>163.60</u>	2216.54
Xuzhou	FLOPs (M)	1024.87	3798.93	120,880.79	15,685.41	**459.19**	5004.17	11,223.26	4997.49	1626.20	33,470.73
	Params (K)	534.66	492.41	99,861.32	1288.17	**153.22**	633.39	1309.58	465.90	<u>252.04</u>	2479.23

The results show that the MGWT has relatively high FLOPs and parameter counts on all three datasets. The convolutional aggregator introduces additional operations when constructing group proxies, while the spatial–spectral dual-stream attention computes spatial and spectral dependencies separately. The three-stage hierarchy further increases the cost while facilitating local and global information modeling. Therefore, the MGWT trades computational efficiency for improved classification performance, and reducing this cost while preserving multi-granularity spatial–spectral modeling remains an important direction for future work.

3.4. Parameter Analysis

3.4.1. Effect of the Number of Groups in the S^2-GPRT Module

The group number determines how many sub-bands are processed by the grouped convolution blocks. To investigate its impact on the performance of the MGWT, ablation experiments were conducted on all three datasets, as summarized in Table 10.

Here, setting the number of groups to one corresponds to standard convolution, which serves to validate the effectiveness of the sub-band splitting strategy. The results indicate that the optimal numbers of groups are four, two, and two for the *PaviaU*, *Houston2013*, and *Xuzhou* datasets, respectively.

Table 10. Effect of different group numbers in S^2-GPRT on three datasets.

Groups	PaviaU	Houston2013	Xuzhou
1	97.72 ± 0.27	91.34 ± 0.61	95.74 ± 0.48
2	97.83 ± 0.42	**92.04 ± 0.34**	**96.79 ± 0.53**
4	**97.83 ± 0.41**	91.84 ± 0.64	96.05 ± 0.95
8	97.79 ± 0.40	91.81 ± 0.58	96.45 ± 0.40
16	97.39 ± 0.39	91.01 ± 0.64	96.52 ± 0.40

3.4.2. Effect of the Number of Encoders in Each Stage

In the three-stage hierarchical feature extraction network, the encoder at each stage is capable of extracting spatial–spectral features of HSIs at different granularities. To investigate the impact of the number of encoders per stage on the performance of the MGWT, the maximum number of encoders in each stage was limited to two, resulting in eight possible combinations, as summarized in Table 11.

Table 11. Effect of the number of Mixformer encoder blocks at each stage on three datasets.

Encoder Layers	PaviaU	Houston2013	Xuzhou
[1, 1, 1]	98.06 ± 0.36	92.08 ± 0.28	96.14 ± 0.71
[2, 1, 1]	97.88 ± 0.46	**92.28 ± 0.37**	96.42 ± 0.48
[1, 2, 1]	97.83 ± 0.41	92.04 ± 0.34	**96.79 ± 0.53**
[1, 1, 2]	**98.45 ± 0.38**	91.94 ± 0.39	96.36 ± 0.63
[2, 2, 1]	97.94 ± 0.44	91.88 ± 0.60	96.17 ± 0.73
[2, 1, 2]	98.10 ± 0.36	92.02 ± 0.33	96.14 ± 0.66
[1, 2, 2]	98.07 ± 0.24	91.95 ± 0.58	96.23 ± 0.93
[2, 2, 2]	97.99 ± 0.23	91.67 ± 0.43	96.14 ± 0.88

The results indicate that the optimal numbers of encoders for the *PaviaU*, *Houston2013*, and *Xuzhou* datasets are [1, 1, 2], [2, 1, 1], and [1, 2, 1], respectively.

3.4.3. Influence of the Patch Size

We evaluated the effect of input patch size on the MGWT across all three datasets. In the experiments, the patch sizes were set to 7 × 7, 9 × 9, 11 × 11, 13 × 13, 15 × 15, 17 × 17, and 19 × 19, and the results are reported in Table 12.

The results indicate that the selected patch sizes for the *PaviaU*, *Houston2013*, and *Xuzhou* datasets are 13 × 13, 9 × 9, and 11 × 11, respectively. On *PaviaU*, 13 × 13 and 15 × 15 attain the same mean OA, and the smaller 13 × 13 setting is selected.

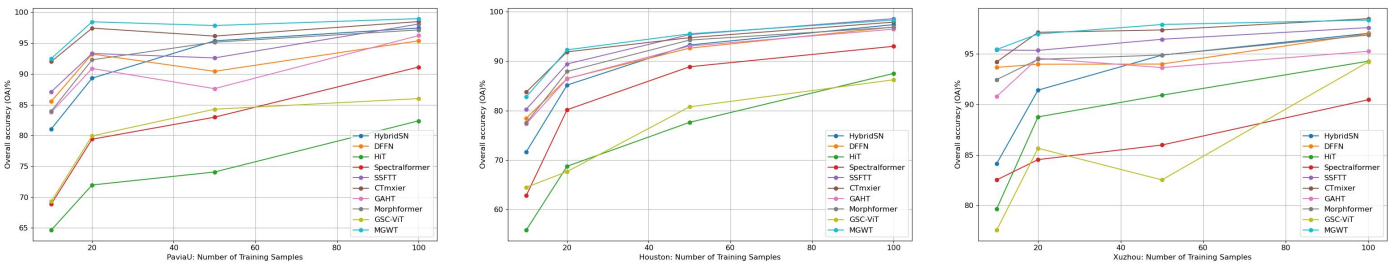
3.4.4. Influence of the Number of Training Samples

To evaluate the adaptability of the model under limited training data, we randomly selected 10, 20, 50, and 100 samples per class from each of the three datasets as the training sets and observed how the model performance varies with the number of training samples,

as shown in Figure 10. The results indicate that, when the number of training samples is small (e.g., 10 samples per class), there are significant differences in classification accuracy among different methods. However, as the amount of training data increases, the performance gap gradually narrows, suggesting that sufficient training data generally helps to improve classification performance. Notably, across different training sample sizes, the MGWT achieves superior performance in most cases, demonstrating its effectiveness in learning spatial–spectral features and obtaining robust classification results.

Table 12. Classification performance of MGWT at different input patch sizes on the three datasets.

Patch Size	PaviaU	Houston2013	Xuzhou
7 × 7	96.17 ± 0.56	91.50 ± 0.37	95.82 ± 0.33
9 × 9	97.21 ± 0.57	**92.28 ± 0.37**	96.79 ± 0.53
11 × 11	97.96 ± 0.47	92.17 ± 0.28	**96.96 ± 0.44**
13 × 13	**98.45 ± 0.38**	91.68 ± 0.41	96.00 ± 0.73
15 × 15	98.45 ± 0.38	91.42 ± 0.50	95.90 ± 0.59
17 × 17	98.28 ± 0.18	90.55 ± 0.50	94.92 ± 0.71
19 × 19	97.83 ± 0.23	89.56 ± 0.54	94.71 ± 0.51



(**a**) (**b**) (**c**)

Figure 10. The influence of the number of training samples on the performance of different methods on the three datasets. (**a**) *PaviaU*, (**b**) *Houston2013*, and (**c**) *Xuzhou*.

4. Discussion

The results across the three datasets show that the MGWT improves overall classification while reducing confusion among several spectrally or visually similar classes and producing cleaner class boundaries. The ablation results indicate that the S^2-GPRT, the multi-granularity aggregator, and spatial–spectral dual-stream attention provide the principal feature-representation capability, whereas the LMTA supplies a smaller but consistent complementary improvement. The complexity comparison also shows that these gains are accompanied by relatively high FLOPs and parameter counts, reflecting a trade-off between classification performance and computational efficiency.

The MGWT has practical limitations when transferred to unseen datasets. Its group number, encoder allocation, and patch size may require dataset-specific selection, while the LMTA relies on four fixed candidate ratios whose suitability may vary across sensors and scenes. This issue can be more pronounced when validation labels are unavailable or limited, and results on three benchmark scenes do not establish robustness under sensor or cross-scene shift. Future work will investigate data-adaptive parameter selection and cross-dataset evaluation using an independent validation protocol.

5. Conclusions

We developed the MGWT, a multi-granularity transformer network for hyperspectral image classification. The MGWT combines the CNN-based S^2-GPRT with a three-stage transformer encoder. By exploiting the physical grouping property of hyperspectral imaging,

the S^2-GPRT uses grouped convolutions for band partitioning and dual-branch hybrid convolutions for multi-scale shallow spatial–spectral feature extraction, resulting in improved local feature modeling. As the core module of the Mixformer encoder, MGA combines a token-level pathway and multi-scale group-proxy pathways through an aggregator, extracts spatial–spectral features via the SpaAS and SpeAS modules, and uses the LMTA to construct and adaptively combine four rank-guided attention branches. Moreover, to capture the spatial–spectral features at different scales, a hierarchical network structure based on the Mixformer encoder is designed. The experimental results on multiple benchmark datasets demonstrate that the MGWT outperforms the existing state-of-the-art methods.

Author Contributions: Conceptualization, Z.X. and S.H.; methodology, Z.X. and W.X.; software, Z.X. and W.X.; validation, A.Z., T.H. and B.Z.; formal analysis, Z.X.; investigation, S.K.B.; resources, S.H.; data curation, W.X.; writing—original draft preparation, Z.X.; writing—review and editing, S.H., T.H. and B.Z.; visualization, A.Z.; supervision, S.H.; project administration, S.H.; funding acquisition, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under grant 42301425, in part by the China Postdoctoral Science Foundation (2023M743299), in part by the "CUG Scholar" Scientific Research Funds at China University of Geosciences (Wuhan) under grant 2022164, in part by the Science and Technology Department of Hubei Province under grant 2025CSA049, in part by the Joint Open Fund of the Research Platforms of School of Computer Science, China University of Geosciences, Wuhan (No. PTLH2024-A-02) and in part by the Open Research Fund of Jiangxi Provincial Engineering Research Center for Territorial Spatial Monitoring and Planning Control under grant JXYJY2025003.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The PaviaU and Houston2013 datasets are publicly available. The Xuzhou dataset is available upon reasonable request from the corresponding author.

Conflicts of Interest: Author Bin Zhou was employed by the company No. 722 Research Institute of China State Shipbuilding Corporation Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ram, B.G.; Oduor, P.; Igathinathane, C.; Howatt, K.; Sun, X. A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects. *Comput. Electron. Agric.* **2024**, *222*, 109037. [CrossRef]
2. Ji, C.; Jilge, M.; Heiden, U.; Stellmes, M.; Feilhauer, H. Sampling Robustness in Gradient Analysis of Urban Material Mixtures. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4500211. [CrossRef]
3. Luo, S.; Yu, J.; Xi, Y.; Liao, X. Aircraft target detection in remote sensing images based on improved YOLOv5. *IEEE Access* **2022**, *10*, 5184–5192. [CrossRef]
4. Huang, S.; Zhao, K.; Chen, H.; Zhang, H. Heterogeneous Data-Based Global-to-Local Cross-Domain Few-Shot Classification Method of Hyperspectral Image. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–19. [CrossRef]
5. Ma, L.; Crawford, M.M.; Tian, J. Local Manifold Learning-Based *k*-Nearest-Neighbor for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4099–4109. [CrossRef]
6. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
7. Zhang, Y.; Cao, G.; Li, X.; Wang, B. Cascaded Random Forest for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1082–1094. [CrossRef]
8. Chen, Y.; Nasrabadi, N.M.; Tran, T.D. Classification for hyperspectral imagery based on sparse representation. In Proceedings of the 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, Reykjavik, Iceland, 14–16 June 2010; pp. 1–4.

9. Sima, H.; Wang, J.; Guo, P.; Sun, J.; Liu, H.; Xu, M.; Zou, Y. Composite Kernel of Mutual Learning on Mid-Level Features for Hyperspectral Image Classification. *IEEE Trans. Cybern.* **2022**, *52*, 12217–12230. [CrossRef]
10. Sadeghi-Tehran, P.; Virlet, N.; Hawkesford, M.J. A neural network method for classification of sunlit and shaded components of wheat canopies in the field using high-resolution hyperspectral imagery. *Remote Sens.* **2021**, *13*, 898. [CrossRef]
11. Huang, S.; Zhang, H.; Xue, J.; Pizurica, A. Heterogeneous Regularization-Based Tensor Subspace Clustering for Hyperspectral Band Selection. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 9259–9273. [CrossRef]
12. Li, M.; Zhang, Q. Local entropy and nonextensivity of networks ensemble. *Commun. Nonlinear Sci. Numer. Simul.* **2026**, *152*, 109369. [CrossRef]
13. Dong, S.; Feng, W.; Quan, Y.; Dauphin, G.; Gao, L.; Xing, M. Deep ensemble CNN method based on sample expansion for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
14. He, L.; Li, J.; Plaza, A.; Li, Y. Discriminative Low-Rank Gabor Filtering for Spectral–Spatial Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 1381–1395. [CrossRef]
15. Liang, L.; Zhang, S.; Li, J. Multiscale DenseNet Meets with Bi-RNN for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5401–5415. [CrossRef]
16. Song, T.; Wang, Y.; Gao, C.; Chen, H.; Li, J. MSLAN: A Two-Branch Multidirectional Spectral–Spatial LSTM Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
17. Hang, R.; Liu, Q.; Hong, D.; Ghamisi, P. Cascaded Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5384–5394. [CrossRef]
18. Yu, Z.; Cui, W. Robust hyperspectral image classification using generative adversarial networks. *Inf. Sci.* **2024**, *666*, 120452. [CrossRef]
19. Liu, Q.; Xiao, L.; Yang, J.; Wei, Z. CNN-enhanced graph convolutional network with pixel-and superpixel-level feature fusion for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 8657–8671. [CrossRef]
20. Zhang, Y.; Wang, X.; Jiang, X.; Zhou, Y. Robust Dual Graph Self-Representation for Unsupervised Hyperspectral Band Selection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
21. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep convolutional neural networks for hyperspectral image classification. *J. Sens.* **2015**, *2015*, 1–12. [CrossRef]
22. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. In *Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*; IEEE: Piscataway, NJ, USA, 2015; pp. 4959–4962.
23. Ghosh, A.; Chidambaram, S. Hyperspectral Image Classification using Archimedes Optimized 3D CNN layers. In *Proceedings of the 2025 5th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*; IEEE: Piscataway, NJ, USA, 2025; pp. 151–157.
24. Roy, S.; Krishna, G.; Dubey, S.; Chaudhuri, B. HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 277–281. [CrossRef]
25. Liu, X.; Ng, A.H.M.; Ge, L.; Lei, F.; Liao, X. Multibranch Fusion: A Multibranch Attention Framework by Combining Graph Convolutional Network and CNN for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [CrossRef]
26. Sun, K.; Dong, F.; Liu, W.; Wu, Q.; Sun, X.; Wang, W. Hyperspectral image classification with spatial-spectral-channel 3D attention and channel attention. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–14. [CrossRef]
27. Moghaddam, S.H.A.; Gazor, S.; Homayouni, S.; Karami, F. Multiscale Deformable DenseNet for Wetland Mapping Using Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5. [CrossRef]
28. Bhatti, U.A.; Yu, Z.; Chanussot, J.; Zeeshan, Z.; Yuan, L.; Luo, W.; Nawaz, S.A.; Bhatti, M.A.; Ain, Q.U.; Mehmood, A. Local Similarity-Based Spatial–Spectral Fusion Hyperspectral Image Classification with Deep CNN and Gabor Filtering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
29. Yu, C.; Han, R.; Song, M.; Liu, C.; Chang, C.I. Feedback Attention-Based Dense CNN for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
30. Sun, H.; Zheng, X.; Lu, X.; Wu, S. Spectral–Spatial Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3232–3245. [CrossRef]
31. Lu, Z.; Xu, B.; Sun, L.; Zhan, T.; Tang, S. 3-D Channel and Spatial Attention Based Multiscale Spatial–Spectral Residual Network for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4311–4324. [CrossRef]

32. Cui, Y.; Luo, L.; Wang, L.; Gao, S.; Zhao, C. CNN enhanced hypergraph attention network for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5. [CrossRef]
33. Huang, S.; Xiao, W.; Chen, H.; Bejo, S.K.; Zhang, H. Hyperspectral image classification based on a locally enhanced transformer network. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–17. [CrossRef]
34. Huang, S.; Zeng, H.; Chen, H.; Zhang, H. Spatial and cluster structural prior-guided subspace clustering for hyperspectral image. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [CrossRef]
35. Zhu, M.; Jiao, L.; Liu, F.; Yang, S.; Wang, J. Residual Spectral–Spatial Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 449–462. [CrossRef]
36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
38. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral Image Classification Using Group-Aware Hierarchical Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
39. Roy, S.; Deria, A.; Shah, C.; Haut, J.; Du, Q.; Plaza, A. Spectral–Spatial Morphological Attention Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
40. Zhao, G.; Ye, Q.; Sun, L.; Wu, Z.; Pan, C.; Jeon, B. Joint Classification of Hyperspectral and LiDAR Data Using a Hierarchical CNN and Transformer. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
41. He, Y.; Tu, B.; Liu, B.; Li, J.; Plaza, A. HSI-MFormer: Integrating mamba and transformer experts for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–16. [CrossRef]
42. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
43. Shi, H.; Zhang, Y.; Cao, G.; Yang, D. MHCFormer: Multiscale Hierarchical Conv-Aided Fourierformer for Hyperspectral Image Classification. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–15. [CrossRef]
44. Jia, S.; Min, Z.; Fu, X. Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion. *Inf. Fusion* **2023**, *96*, 117–129. [CrossRef]
45. Fang, Y.; Sun, L.; Zheng, Y.; Wu, Z. Deformable Convolution-Enhanced Hierarchical Transformer with Spectral-Spatial Cluster Attention for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2025**, *34*, 701–716. [CrossRef]
46. Jiang, M.; Su, Y.; Gao, L.; Plaza, A.; Zhao, X.L.; Sun, X.; Liu, G. GraphGST: Graph generative structure-aware transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef]
47. Sun, L.; Zhang, H.; Zheng, Y.; Wu, Z.; Ye, Z.; Zhao, H. MASSFormer: Memory-Augmented Spectral-Spatial Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [CrossRef]
48. Zhao, Z.; Xu, X.; Li, S.; Plaza, A. Hyperspectral image classification using groupwise separable convolutional vision transformer network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [CrossRef]
49. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 367–376.
50. Yang, X.; Cao, W.; Tang, D.; Zhou, Y.; Lu, Y. ACTN: Adaptive Coupling Transformer Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. [CrossRef]
51. Peng, Y.; Zhang, Y.; Tu, B.; Li, Q.; Li, W. Spatial–Spectral Transformer with Cross-Attention for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5537415. [CrossRef]
52. He, Y.; Tu, B.; Liu, B.; Chen, Y.; Li, J.; Plaza, A. Hybrid Multiscale Spatial–Spectral Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–18. [CrossRef]
53. Zou, J.; He, W.; Zhang, H. LESSFormer: Local-Enhanced Spectral-Spatial Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
54. Wang, X.; Tan, K.; Du, Q.; Chen, Y.; Du, P. Caps-TripleGAN: GAN-Assisted CapsNet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7232–7245. [CrossRef]
55. Song, W.; Li, S.; Fang, L.; Lu, T. Hyperspectral Image Classification with Deep Feature Fusion Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3173–3184. [CrossRef]
56. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]

57. Yang, X.; Cao, W.; Lu, Y.; Zhou, Y. Hyperspectral image transformer classification networks. *IEEE Trans*. *Geosci*. *Remote Sens*. **2022**, *60*, 1–15. [CrossRef]
58. Zhang, J.; Meng, Z.; Zhao, F.; Liu, H.; Chang, Z. Convolution transformer mixer for hyperspectral image classification. *IEEE Geosci*. *Remote Sens*. *Lett*. **2022**, *19*, 1–5. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.