remote sensing

MDPI

Article

Dual-Branch Spatial–Spectral Transformer with Spatial Feature Enhancement for Hyperspectral Image Classification

Liguo Wang, Teng Wen and Danfeng Liu *

College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China; wangliguo@hrbeu.edu.cn (L.W.); 202311051005@stu.dlnu.edu.cn (T.W.)
* Correspondence: liudanfeng@dlnu.edu.cn

Highlights

What are the main findings?

- We design a dual-branch spatial–spectral transformer for hyperspectral image classification. A deformable spatial feature fusion module captures irregular ground-object spatial structures, and a global spatial semantic aggregation module improves the discrimination of spatial features.
- The spectral branch combines Transformer and a convolutional feed-forward network. It captures global inter-band spectral dependencies, strengthens local spectral correlations, and achieves competitive results on four public hyperspectral datasets.

What are the implications of the main findings?

- The grouped deformable convolution and dual-branch transformer scheme provides an effective solution for modeling complex irregular spatial patterns in hyperspectral imagery.
- The spatial–enhanced dual-branch architecture can serve as a reference for other remote-sensing interpretation tasks requiring joint spatial–spectral feature learning.

Abstract

Hyperspectral image classification is an important research topic in the field of remote sensing. Its core objective is to accurately identify different land-cover categories by jointly exploiting the spatial information and spectral characteristics of pixels. However, the complex and diverse spatial structures of ground objects in hyperspectral images make it challenging for existing methods to fully capture intricate spatial patterns. To address this issue, this paper proposes a dual-branch spatial–spectral transformer with spatial feature enhancement under a spatial–spectral joint representation learning framework. Specifically, in the spatial branch, a deformable spatial feature fusion module is introduced, which combines grouped deformable convolution with Transformer to achieve fine-grained modeling of irregular spatial structures. Subsequently, a global spatial semantic aggregation module is designed to further integrate spatial features and enhance their discriminative capability. In the spectral branch, Transformer is employed to model long-range spectral dependencies, while a convolutional feed-forward network is introduced to strengthen the representation ability of critical spectral information. Experimental results demonstrate that the proposed method achieves competitive classification performance on four benchmark datasets.

Keywords: hyperspectral image classification; grouped deformable convolution; Transformer

Check for updates

Academic Editor: Akira Iwasaki

Received: 7 September 2026
Revised: 16 October 2026
Accepted: 21 October 2026
Published: 26 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.

1. Introduction

With the continuous advancement of sensor technology, acquired hyperspectral images have progressively become sharper and more detailed. Hyperspectral imagery is intrinsically a three-dimensional data cube composed of two spatial dimensions and one spectral dimension, thereby encompassing rich feature information that provides robust discriminative capability for identifying diverse land-cover and land-use classes. Consequently, it has been widely adopted in numerous remote sensing applications, including precision agriculture, geological exploration, and urban planning [1–5].

Hyperspectral image (HSI) classification constitutes one of the core research areas in hyperspectral remote sensing, aiming to assign each pixel an appropriate land-cover class label by jointly exploiting spectral signatures and spatial contextual information—thereby enabling automated land-cover identification. In the early stage of HSI classification research, a wide array of conventional machine learning methods emerged, including K-nearest neighbors [6] (KNNs), support vector machines [7] (SVMs), and random forests [8,9] (RFs). Moreover, to mitigate the "curse of dimensionality" (often referred to as the Hughes phenomenon) inherent in high-dimensional hyperspectral data, researchers proposed several dimensionality reduction techniques—such as principal component analysis [10] (PCA), independent component analysis [11] (ICA), and linear discriminant analysis [12] (LDA). These methods effectively suppress spectral redundancy while preserving discriminative spectral information, thereby significantly enhancing classification performance.

To effectively integrate spatial and spectral information and thereby enhance classification performance, Ye et al. proposed a Gabor feature-based approach combined with decision fusion [13]. Their method systematically incorporates band grouping, spatial feature extraction, feature dimensionality reduction, and classifier-level decision fusion—achieving efficient synergy between spatial and spectral information. Zhang et al. introduced a Gabor filter-based multi-scale dense network [14], which employs 3D-Gabor filters to extract spatial texture features from hyperspectral images and leverages a multi-scale dense architecture to capture spatial–spectral features across multiple orientations. Morphology-based methods [15] utilize morphological opening and closing operations to extract diverse structural features, thereby significantly improving classification accuracy. Chen et al. proposed a sparse representation (SR)-based classification framework [16], in which contextual information is further incorporated through the introduction of a smoothness constraint and the construction of a joint sparse model.

With the continuous advancement of deep learning techniques, their application in computer vision and related fields has become increasingly widespread [17]. Compared with traditional machine learning approaches, deep learning methods offer the distinct advantage of automatically capturing intrinsic spatial–spectral correlations within hyperspectral images, thereby eliminating the need for handcrafted feature engineering. Consequently, deep learning has been widely adopted for hyperspectral image classification [18,19].

Convolutional neural networks (CNNs) have gradually become the mainstream technique in classification tasks owing to their advantages in local feature extraction and weight sharing. Hu et al. applied a CNN architecture with five weight layers to hyperspectral image classification [20]. However, one-dimensional convolution operates solely along the spectral dimension and thus captures only spectral features. To address this limitation, Yang et al. proposed a dual-branch CNN architecture [21], wherein one branch extracts spatial features while the other extracts spectral features; these two complementary feature streams are subsequently fused via fully connected layers to generate higher-level joint spectral–spatial representations. This design enables synergistic integration of spatial and spectral information, thereby enhancing overall classification performance. To further

enable joint modeling of spatial and spectral features and learn deeper, more discriminative joint representations, Sun et al. introduced three-dimensional convolution into the classification framework and proposed a spectral–spatial attention network [22] (SSAN). Specifically, the SSAN first constructs a foundational spectral–spatial network composed of dedicated spectral and spatial modules and then integrates an attention mechanism to suppress the adverse influence of noisy or irrelevant pixels—thus enabling extraction of more discriminative spectral–spatial features. Chen et al. proposed a novel approach [23] that leverages 3D-CNNs to simultaneously extract spatial features, spectral features, and joint spatial–spectral features from hyperspectral images. Moreover, regularization strategies are incorporated to mitigate overfitting, achieving high-accuracy classification results. Recognizing that fixed local receptive fields in conventional CNNs constrain classification performance, Roy et al. proposed the A2S2K-ResNet method [24]. This method employs an enhanced 3D residual block to jointly extract spatial–spectral features and incorporates a feature recalibration mechanism to achieve end-to-end hyperspectral image classification. Cai et al. introduced the dual-stream cross-domain network [25] (DSCD), which, in its second stage, synergistically combines the respective strengths of 2D and 3D convolutions to comprehensively exploit joint spatial–spectral features. Zhu et al. proposed a residual spectral–spatial attention network [26], which utilizes dedicated spatial and spectral attention modules to adaptively weight informative features and thereby efficiently extract discriminative spectral–spatial representations. Furthermore, the attention mechanism is embedded within residual blocks, effectively alleviating overfitting risks.

An increasing number of scholars have recognized that conventionally handcrafted convolutional neural networks (CNNs) struggle to simultaneously achieve effective multi-dimensional information extraction from hyperspectral data and high computational efficiency. To address this challenge, Paoletti et al. proposed AAtt-CNN—a novel automatic attention-based CNN framework for hyperspectral image classification [27]. This method integrates both a one-dimensional spectral classifier and a three-dimensional spatial–spectral classifier, enabling complementary exploitation of hyperspectral data across distinct dimensional domains. Liu et al. introduced a hybrid classification framework that synergistically combines fast dynamic graph convolution with CNNs [28]. Specifically, the framework leverages flattened and pooled feature representations from convolutional layers to comprehensively capture spatial–spectral correlations within the hyperspectral cube; concurrently, a dynamic graph convolution module is employed to model intrinsic structural relationships among image patches. Finally, the framework fuses three complementary feature streams to produce the final classification output.

In addition to the CNN-based methods mentioned earlier, several other advanced models have been proposed for hyperspectral image classification. Zhao et al. introduced a superpixel-guided deformable convolutional network for hyperspectral classification [29], which leverages superpixels as structural priors to guide convolutional sampling, thereby achieving high-accuracy classification of hyperspectral images. Recently, Zhuang et al. proposed a frequency domain iterative clustering superpixel segmentation method that preserves boundaries [30]. By evaluating image complexity through information assessment during the seed initialization stage, they provided a new approach for region partitioning incorporating local details. These studies demonstrate the importance of considering regional organization and boundary structures in spatial representations, offering valuable insights for modeling complex land-cover structures in hyperspectral images.

Mou et al. pioneered the application of recurrent neural network (RNN) architectures to hyperspectral image classification, proposing a deep recurrent neural network (DRNN) framework [31]. Their approach introduces a novel parametric rectified tanh (PReTanh) activation function and accordingly revises the gated recurrent unit (GRU) architecture.

This modification enables efficient processing of hyperspectral data while significantly reducing the model's parameter count. Paoletti et al. proposed a multiple attention-guided capsule network [32], which synergistically integrates the spatial modeling capability of capsule networks with multi-level attention mechanisms. This integration not only alleviates the computational burden associated with parameter-intensive models but also enhances the discriminability of learned features. Feng et al. introduced a novel multiclass spatial–spectral generative adversarial network (MSGAN) for hyperspectral image classification [33]. The method employs two complementary generators—one dedicated to synthesizing spatially coherent samples and the other to generating spectrally faithful samples—while a shared discriminator extracts joint spatial–spectral features and outputs multiclass classification probabilities. Through adversarial training between the generators and the discriminator, high-fidelity synthetic samples are leveraged to progressively improve the discriminator's classification performance. Wan et al. proposed an enhanced graph convolutional network (GCN)-based classification method [34]. Their approach utilizes region-induced dynamic graphs and performs continuous graph convolution to capture long-range contextual dependencies among pixels. Crucially, it adaptively learns both edge weights and graph connectivity during training, enabling iterative refinement of the graph structure to better encode inter-pixel contextual relationships.

Although CNNs exhibit certain advantages in extracting local features, their ability to capture long-range dependencies is inherently limited by the constrained receptive field of convolutional kernels. In 2017, the Google research team introduced Transformer [35]—a novel model architecture that leverages the self-attention (SA) mechanism to compute pairwise relevance scores among all input elements, thereby enabling global contextual modeling without distance-based constraints. In 2020, the same team extended the transformer architecture to computer vision, proposing the Vision Transformer [36] (ViT), which achieved remarkable performance on various benchmark tasks [37,38]. Subsequently, researchers adapted ViT to hyperspectral image classification, offering a promising new paradigm for achieving high-accuracy classification.

Hong et al. approached hyperspectral image classification from the perspective of spectral sequence modeling and proposed the SpectralFormer—a spectral transformer-based method [39]. This method leverages the transformer architecture to learn discriminative spectral features from grouped adjacent spectral bands and incorporates cross-layer skip connections to enhance feature propagation, thereby mitigating feature degradation commonly observed in deep networks. Liao et al. developed an attention-only spatial–spectral network for hyperspectral image classification, built entirely upon the visual transformer framework [40]. Their architecture features two dedicated, decoupled attention modules—namely, a spatial attention module and a spectral attention module—each designed to explicitly model long-range contextual dependencies among pixels in the spatial and spectral domains, respectively. Peng et al. introduced the cross-attention-based spatial–spectral transformer (CASST) [41], a dual-branch transformer architecture. The spatial branch captures fine-grained spatial structures, while the spectral branch processes spectral sequences to extract spectral features and model inter-band dependencies. Crucially, CASST integrates a cross-attention module with weight-sharing mechanisms between the two branches, enabling effective bidirectional interaction and joint spatial–spectral feature learning.

Although transformers outperform CNNs in capturing long-range dependencies, CNNs excel at spatial feature extraction. Consequently, several researchers have begun exploring hybrid architectures that integrate convolutional and transformer-based components. He et al. proposed the spatial–spectral transformer [42] (SST), which employs a VGGNet-like convolutional network to extract spatial features and a densely connected

transformer to model spectral characteristics; the two types of features are then fused to enrich the overall representation. SST empirically validates the effectiveness of synergistic CNN–transformer modeling and lays a foundational framework for subsequent research. Sun et al. introduced a lightweight classification architecture—the spectral–spatial feature tokenization transformer [43] (SSFTT). Its key innovation lies in using a simple convolutional module for shallow feature extraction, followed by a Gaussian-weighted tokenization module and a transformer encoder for high-level feature learning. This design achieves superior classification accuracy while maintaining a significantly reduced parameter count. Inspired by grouped convolutions, Mei et al. proposed the group-aware hierarchical transformer [44] (GAHT). GAHT incorporates a grouped pixel embedding module to capture localized spatial–spectral features, thereby mitigating feature dispersion inherent in multi-head self-attention. Furthermore, its hierarchical multi-layer architecture enables comprehensive modeling of multi-scale features, enhancing representational capacity. To fully exploit the intrinsic structural and spectral properties of hyperspectral data, Roy et al. developed a classification method that integrates morphological operations with transformer-based attention mechanisms [45]. Specifically, morphological convolution is embedded within the attention mechanism to strengthen inter-pixel structural and shape-aware information exchange, thereby enabling more effective joint exploitation of spatial–spectral features and improved classification performance.

To fully leverage the advantages of combining convolutional neural networks with transformers and enhance the representation capability of local-global features, Zhang et al. proposed a convolutional transformer mixer model [46]. This method first extracts local spatial–spectral features using parallel residual blocks and then explores both local and global features through a dual-branch CNN–transformer architecture. Meanwhile, convolutions are embedded into the multi-head self-attention mechanism to construct a local–global MHSA module, enabling more effective fusion of local and global information. To address the limitation of standard transformers in adequately capturing the fine-grained local spatial–spectral characteristics inherent in HSI data, Zhao et al. introduced the groupwise separable convolutional vision transformer [47] (GSC-ViT). This architecture leverages a groupwise separable convolutional module to explicitly model local spatial–spectral patterns and incorporates a groupwise separable self-attention mechanism to achieve synergistic modeling of local and global features. Additionally, Zhao et al. proposed a dual-branch spectral–spatial attention network [48], which captures local–global spatial features through spatial self-attention and scale information extraction modules, while leveraging transformers to model spectral characteristics, achieving high-precision classification. Arshad et al. proposed a lightweight hyperspectral image classification method by synergistically integrating convolutional neural networks (CNNs) and transformers [49]. This approach effectively combines the complementary strengths of transformers and CNNs, enabling simultaneous and efficient capture of both local spatial–spectral features and long-range hierarchical dependencies.

Recent studies have further investigated spectral–spatial feature interaction, multi-scale dependency modeling, and model lightweighting. Zhang et al. [50] proposed the spectral–spatial dual-branch fusion transformer (S2DBFT) for hyperspectral image classification, which employs one-dimensional and two-dimensional convolutions to extract shallow spectral and spatial features, respectively, and enhances cross-modal interaction between these features through weighted fusion and multi-head spectral–spatial self-attention. Wang et al. [51] proposed the mid-range convolutional modulated transformer network (MCMTN), which extracts short-range spatial and spectral features through a lightweight spatial–spectral module and effectively captures mid-range spatial dependencies in hyperspectral images by incorporating a mid-range convolutional modulation mechanism.

Zhong et al. [52] proposed the lightweight convolution-transformer network (LCTNet), which integrates spectral feature extraction and dimensionality reduction, spatial–spectral feature tokenization, and dynamic sparse transformers to enhance key spatial–spectral features while adjusting attention sparsity to reduce redundant computations.

Among the aforementioned methods, the effective integration of convolutional neural networks and transformers has enhanced hyperspectral image classification performance. However, existing approaches remain insufficient in modeling spatial features during joint spatial–spectral representation learning. First, conventional convolution operations typically rely on fixed sampling locations to extract local spatial features. However, objects in hyperspectral images often exhibit irregular geometries and complex spatial distributions, rendering fixed-sampling strategies ill-suited to capturing such heterogeneous structural patterns—potentially resulting in suboptimal extraction of fine-grained local spatial details. Although deformable convolutions have been explored in some recent studies for hyperspectral image classification, the effective integration of adaptive local spatial modeling with global contextual representation learning remains an open challenge warranting further investigation. Second, relying solely on local spatial feature extraction makes it difficult to obtain sufficiently discriminative feature representations. Features from different spatial locations need to be further integrated to capture higher-level spatial semantic information. Therefore, it is necessary to further strengthen spatial feature learning to enhance the representational capability of spatial features. Moreover, the spectral information contained in hyperspectral images exhibits not only long-range dependencies across distant bands but also strong local correlations among adjacent bands. Consequently, effective spectral feature extraction must jointly model both global spectral dependencies and local spectral correlations to yield a more comprehensive and discriminative representation of spectral characteristics.

To address the aforementioned issues, this paper proposes a dual-branch spatial–spectral transformer with spatial feature enhancement. In the spatial branch, grouped deformable convolution is combined with Transformer to enhance the modeling capability for irregular local spatial structures and spatial contextual relationships; subsequently, global spatial semantic information is further aggregated using grouped convolution and a transformer incorporating classification tokens. In the spectral branch, self-attention mechanisms are integrated with convolutional feed-forward networks to simultaneously capture global spectral dependencies and local spectral correlations, enabling complementary learning between spatial and spectral features. The main contributions of this paper are as follows:

(1) The deformable spatial feature fusion module enhances the model's capacity to represent complex and irregular spatial structures by introducing learnable sampling offsets to adapt the fixed sampling locations of conventional convolution, thereby enabling more effective modeling of irregular geographic object shapes.
(2) The global spatial semantic aggregation module enhances the semantic representation of features through grouped convolution and further integrates spatial semantics transformer architecture, thereby improving the discriminative capability of spatial features.
(3) In the spectral branch, a self-attention mechanism is employed to model the spectral vector of the central pixel, thereby capturing global inter-band correlations across the entire spectrum; this is further enhanced by a convolutional feed-forward network to strengthen the representation capability of salient spectral features.

2. Methodology

2.1. Overview

The overall architecture of the proposed method is illustrated in Figure 1, comprising four key components: hyperspectral data preprocessing, a deformable spatial feature fusion module within the spatial branch, a global spatial semantic aggregation module, and a spectral transformer module within the spectral branch. Hyperspectral data are naturally represented as a three-dimensional cube, denoted as $X \in \mathbb{R}^{H\times W\times B}$, serving as the initial input to the method, where H and W denote the height and width dimensions, respectively, and B represents the number of spectral bands. Subsequently, principal component analysis (PCA) is applied for dimensionality reduction, yielding the transformed tensor $X_{pca} \in \mathbb{R}^{H\times W\times b}$.

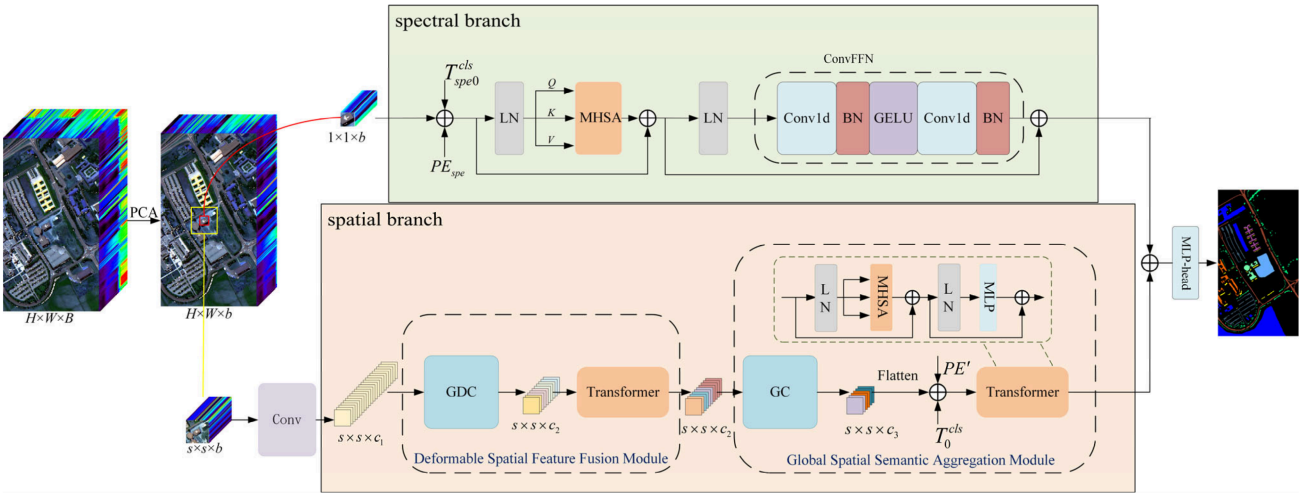


Figure 1. Overall framework diagram.

From X_{pca}, three-dimensional patches $X_{patch} \in \mathbb{R}^{s\times s\times b}$ are extracted via a sliding-window strategy to serve as input to the spatial branch. Let (i, j) denote the spatial coordinates of the central pixel of a given patch, subject to the constraints $0 \le i < H$ and $0 \le j < W$. Defining the window half-length as $r = (s-1)/2$, the patch spans rows from $i-r$ to $i+r$ and columns from $j-r$ to $j+r$. To ensure consistent patch extraction for all central pixels—including those near image boundaries—zero-padding is applied to regions extending beyond the image borders. Concurrently, the spectral vector corresponding to the central pixel of each patch, denoted as $X_{spe} \in \mathbb{R}^{1\times 1\times b}$, is extracted and fed into the spectral branch.

2.2. Deformable Spatial Feature Fusion Module

The deformable spatial feature fusion module (DSFFM), serving as the first module in the spatial branch, provides foundational features for subsequent refinement stages. As illustrated in Figure 1, this module integrates grouped deformable convolution (GDC) and a transformer architecture. Specifically, the input feature maps are partitioned along the channel dimension via GDC, and within each group, spatial sampling offsets are learned independently, enabling fine-grained modeling of spatial structures. Subsequently, the transformer is employed to model the sequence of spatial features, thereby effectively enhancing contextual awareness.

Before performing grouped deformable convolution, the input features are first adjusted in channels via a convolutional layer to obtain the feature map $X^{DSFFM} \in \mathbb{R}^{s\times s\times C_{in}^1}$. Subsequently, this feature map is uniformly divided along the channel dimension into G groups, with each group containing C_{in}^1/G channels. Each group independently per-

forms deformable convolution using its corresponding feature channels to extract distinct spatial features.

Specifically, a convolutional layer first generates an offset feature map $\Delta \in \mathbb{R}^{s\times s\times(2GK^2)}$, where G denotes the number of groups and K represents the spatial size of the convolution kernel. The generated offsets are also partitioned along the channel dimension into G groups, with each group containing $2K^2$ channels—corresponding to horizontal and vertical offset values for K^2 sampling locations of the convolution kernel. Meanwhile, the convolution weights are similarly divided into G groups.

Within each group, deformable convolution is computed using the corresponding input features, sampling offsets, and convolution weights to obtain the respective spatial features. Finally, the output features from all groups are concatenated along the channel dimension to form the final output feature map. The processes of grouping for features, offsets, and convolution weights are respectively illustrated in Equations (1)–(3). Let the range of group index g be $[0, 1, \ldots, G-1]$. Figure 2 demonstrates the grouped deformable convolution operation when $G = 2$.

$$X_g^{DSFFM} = X^{DSFFM}\left[g\frac{G_{in}^1}{G} : (g+1)\frac{G_{in}^1}{G}, :, :\right] \tag{1}$$

$$\Delta_g = \Delta\left[2gK^2 : 2(g+1)K^2, :, :\right] \tag{2}$$

$$W_g = W\left[g\frac{G_{out}^1}{G} : (g+1)\frac{G_{out}^1}{G}, :, :, :\right] \tag{3}$$

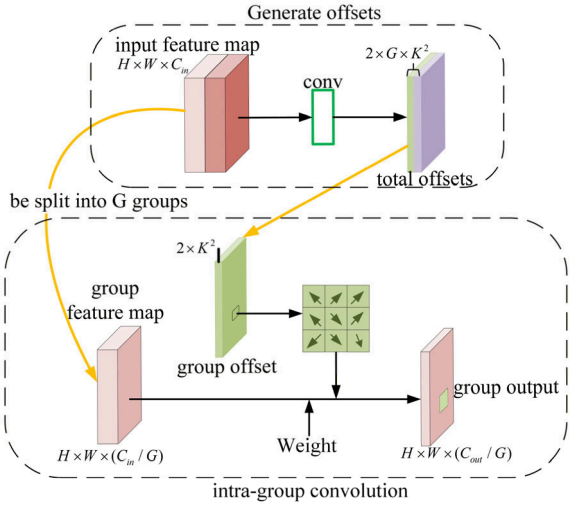


Figure 2. Grouped deformable convolution operation schematic.

First, dynamic offsets are applied to each sampling location $P_k\left(k = 1, 2, \ldots, K^2\right)$ of the standard convolutional kernel. The dynamically offset sampling location for the g-th group is thus given by $P_{g,k} = P_k + \Delta_{g,k}$, where $\Delta_{g,k}$ denotes the corresponding two-dimensional spatial offset vector. Subsequently, within each group, the deformed sampling locations—determined jointly by the group-specific offsets and the convolutional weights—are used to sample the input feature map via bilinear interpolation $f(\cdot)$, yielding the group-wise output feature $Y_g^{DSFFM}(p)$, where p denotes the spatial coordinate on the feature map. Finally, the outputs from all groups are concatenated along the channel dimension to produce

the final output feature map $Y^{DSFFM} \in \mathbb{R}^{s\times s\times C^1_{out}}$. This process is formally described in Equations (4) and (5):

$$Y_g^{DSFFM}(p) = \sum_{k=1}^{K^2} W_g[K]\cdot f\left(X_g^{DSFFM}\left(P_{g,k}\right)\right) \tag{4}$$

$$Y^{DSFFM} = \Phi\left(BN\left(Concat\left(Y_0^{DSFFM}, Y_1^{DSFFM}, \ldots, Y_{G-1}^{DSFFM}\right)\right)\right) \tag{5}$$

Herein, *Concat* denotes the concatenation operation, while BN and Φ represent batch normalization and the activation function, respectively.

By performing grouped partitioning on the input feature maps, offset fields, and convolutional weights, the proposed method ensures that each group maintains an independent set of offset sampling points. Moreover, different groups operate independently on disjoint subsets of features, with their respective parameters fully decoupled. This design enhances the model's capacity to represent diverse spatial structures and improves its generalization performance. However, standard convolutional operations alone are inherently limited in modeling long-range spatial dependencies. To address this limitation, a transformer module is introduced after the grouped deformable convolution (GDC), as illustrated in Figure 3.

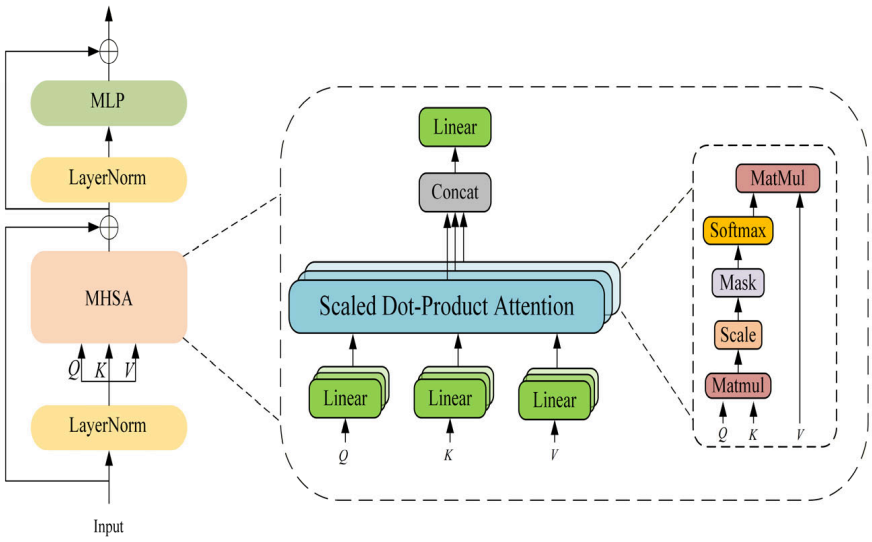


Figure 3. Schematic diagram of the transformer architecture.

The self-attention mechanism takes three inputs: the query vector Q, the key vector K, and the value vector V. It computes attention scores via dot products between Q and K, thereby quantifying pairwise feature correlations. The resulting correlation matrix is then applied to weight the value vectors V, enabling adaptive reweighting of features based on their contextual relevance. This process yields output features with dynamically calibrated attention across the sequence. The input sequence is formally defined in Equation (6). Equations (7) and (8) formulate the computational process of the multi-head self-attention mechanism:

$$T_{in} = [T_1, T_2, \ldots, T_w] + PE \tag{6}$$

$$SA = Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

$$MHSA = Concat(SA_1, SA_2, \ldots, SA_h)W_o \tag{8}$$

Among these, PE denotes the positional encoding; d_k represents the feature dimensionality of the keys; h denotes the number of attention heads; and W_o is the linear projection matrix applied at the output layer.

After passing through the multi-head self-attention mechanism, the features are fed into a feed-forward network (FFN)—comprising two fully connected layers with a GELU activation function applied between them—to perform additional nonlinear transformations on the attention-derived representations, thereby enhancing the model's expressive capacity. Upon completion of the encoding process, the output sequence T_{out} is remapped back to the original spatial structure, preserving identical spatial dimensions as the input features. DSFFM leverages the complementary strengths of grouped deformable convolution and the transformer architecture to achieve fine-grained spatial structural modeling, thereby yielding spatial features with enhanced representational capacity.

2.3. Global Spatial Semantic Aggregation Module

Following DSFFM processing, the model's capability to represent complex spatial structures is enhanced. To further improve the semantic representation of spatial features, this paper introduces a global spatial semantic aggregation module (GSSAM) subsequent to DSFFM. This module performs additional integration of the spatial features output from the preceding stage. As illustrated in Figure 1, GSSAM comprises a grouped convolution (GC) layer and a transformer encoder. Specifically, GSSAM first employs grouped convolution to enrich the semantic expressiveness of features; it then leverages the transformer encoder to model long-range spatial relationships among features, thereby achieving deeper semantic aggregation and enhancing the discriminative power of spatial representations.

Specifically, the input features are first reorganized and integrated along the channel dimension via grouped convolution. The input feature map $X^{GSSAM} \in \mathbb{R}^{s\times s\times C_{in}^2}$ is evenly partitioned into g sub-features X_g^{GSSAM}, where g ranges over $[0, 1, \ldots, G-1]$. Each subgroup undergoes convolution using its dedicated weight matrix. Subsequently, the resulting feature maps are concatenated along the channel dimension to yield the output feature map $Y^{GSSAM} \in \mathbb{R}^{s\times s\times C_{out}^2}$.

Group convolution further enhances the representation of spatial–semantic features while maintaining computational efficiency, thereby providing more robust feature representations for subsequent global relational modeling by the transformer. Subsequently, the features are fed into the transformer encoder in the form of spatial sequences, as illustrated in Equation (9):

$$T'_{in} = \left[T_0^{cls}, T'_1, \ldots, T'_w\right] + PE' \tag{9}$$

Among them, T_0^{cls} denotes a learnable classification token, and PE' represents the positional encoding.

GSSAM likewise employs the standard self-attention mechanism and feed-forward network. The key distinction lies in the incorporation of a learnable classification token into the input sequence. After spatial features are enhanced via grouped convolution, they are fed into the transformer encoder, enabling further integration of feature relationships across distinct spatial locations. During this process, the classification token effectively aggregates global spatial information, thereby yielding a high-level semantic representation that serves as the output feature of the spatial branch.

2.4. Spectral Transformer Module

Hyperspectral image data contain rich spectral features; effective modeling of these spectral features enhances feature discriminability, thereby improving the classification accuracy of models. Unlike convolutional neural networks, transformers can directly model global dependencies across the spectral dimension and capture complex nonlinear

relationships among spectral bands. This capability enables them to emphasize salient spectral features while simultaneously suppressing noise and irrelevant information within the spectral sequence, further boosting classification performance. Accordingly, this paper proposes a spectral transformer module (SpeFormer) for integration into the classification task, whose architecture is illustrated in Figure 1.

This module takes the spectral vector of the central pixel within each three-dimensional data block of the feature map as input, thereby preserving critical spectral information while reducing computational complexity. The input sequence is formulated in Equation (10). The transformer computation follows a standard encoder architecture:

$$T_{spe} = \left[T_{spe0}^{cls}, T_{spe1}, \ldots, T_{speb}\right] + PE_{spe} \tag{10}$$

In this equation, T_{spe0}^{cls} denotes the learnable token for the spectral branch and PE_{spe} represents the positional encoding.

The multi-layer perceptron (MLP), which constitutes the standard feed-forward network, primarily performs nonlinear feature mapping. However, it exhibits limited capability in extracting local features and tends to overlook local correlations inherent in spectral sequences. Moreover, its fully connected layers entail a large number of parameters, collectively constraining model performance to some extent. To better capture local relationships among adjacent spectral bands, this paper replaces the conventional feed-forward network with a convolutional feed-forward network (ConvFFN). As illustrated in Figure 1, the ConvFFN comprises two one-dimensional convolutional layers, followed by batch normalization and a GELU activation function. It integrates the input features with the convolved features via residual connections. This architecture enhances local feature interactions across adjacent spectral bands while preserving the integrity of the original feature representation. As shown in Equation (11), the ConvFFN is formulated as follows:

$$ConvFFN = f_2(GELU(f_1(MHSA))) \tag{11}$$

Herein, f_1 and f_2 denote two composite operations, each consisting of a one-dimensional convolution followed by batch normalization.

This design preserves the transformer's capability to capture global dependencies while simultaneously enhancing local feature modeling among adjacent spectral bands. The output features from the spectral branch are concatenated with those from the spatial branch along the channel dimension and subsequently fed into a classifier for final classification. The classifier adopted in this work consists of an MLP head comprising layer normalization followed by fully connected layers. The output dimension of the final fully connected layer matches the total number of target classes; thus, the class associated with the highest score is designated as the predicted label for the given pixel.

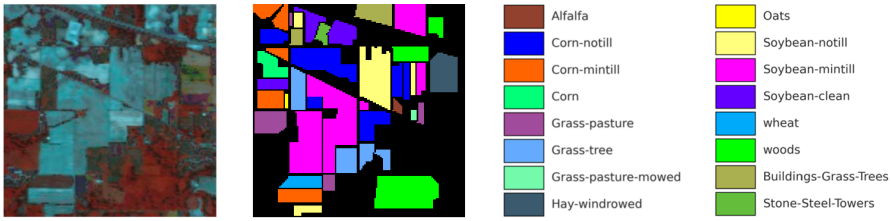
3. Experiments

3.1. Dataset Introduction

To validate the effectiveness of the proposed method in modeling complex spatial structures, this paper conducts experimental analysis on four representative hyperspectral image datasets: the Indian Pines dataset, the Pavia University dataset, the Xuzhou dataset, and the WHU-Hi-LongKou dataset.

The Indian Pines dataset captures a scene located in the northwestern region of Indiana, USA, encompassing both pine forest and agricultural land. Acquired by the airborne visible/infrared imaging spectrometer (AVIRIS) produced in the United States, the dataset features a spatial resolution of approximately 20 m and comprises a 145 × 145-pixel image.

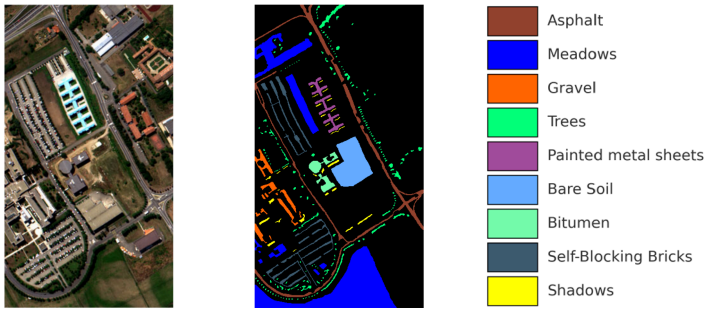
It covers 16 distinct land-cover classes, predominantly agricultural crops. The original hyperspectral image contains 220 contiguous spectral bands; after removing bands affected by atmospheric absorption and sensor noise, 200 bands remain usable. Figure 4 presents a false-color composite image of the Indian Pines dataset alongside its corresponding ground-truth label map.



(**a**) false-color map (**b**) ground truth map

Figure 4. The Indian Pines dataset.

The Pavia University dataset captures scenes encompassing the campus of the University of Pavia and its surrounding urban area. Acquired in 2003 using a reflective optics spectral imaging system (ROSIS), the dataset features a spatial resolution of approximately 1.3 m and an image size of 610 × 340 pixels. It covers nine representative land-cover classes, predominantly including urban buildings, roads, and vegetation. The original hyperspectral image comprises 115 spectral bands; after removing noisy and uncalibrated bands, 103 spectrally informative bands are retained for experimental analysis. Figure 5 presents the false-color composite image of the Pavia University dataset alongside its corresponding ground-truth label map.

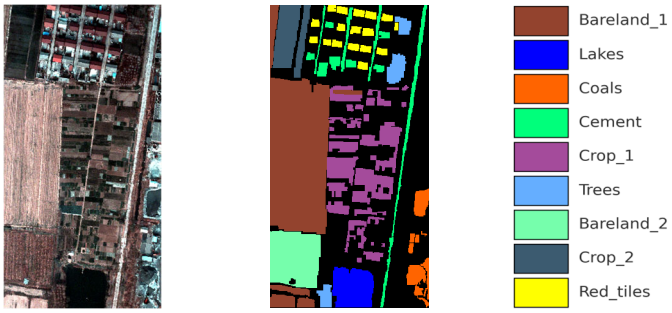


(**a**) false-color map (**b**) ground truth map

Figure 5. The Pavia University dataset.

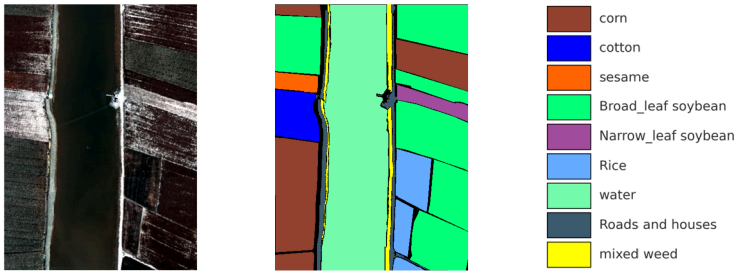
The Xuzhou dataset was acquired using an airborne HYSPEX hyperspectral camera over a suburban mining area in Xuzhou City, Jiangsu Province, China. It features a high spatial resolution of 0.73 m and comprises imagery of 500 × 260 pixels in dimension. The dataset encompasses nine distinct land-cover classes, primarily including coal mines, croplands, and vegetation. The original hyperspectral image contains 436 spectral bands. Figure 6 presents the false-color image of the Xuzhou dataset alongside its corresponding ground-truth label map.

The WHU-Hi-LongKou dataset was acquired in Longkou Town, Hubei Province, China. It represents a relatively simple agricultural scene comprising nine distinct crop categories. The dataset features a spatial resolution of 0.463 m, an image size of 550 × 400 pixels, and 270 effective spectral bands. Figure 7 presents a false-color image of the WHU-Hi-LongKou dataset alongside its corresponding ground-truth label map.



(**a**) false-color map (**b**) ground truth map

Figure 6. The Xuzhou dataset.



(**a**) false-color map (**b**) ground truth map

Figure 7. The WHU-Hi-LongKou dataset.

3.2. Experimental Setup

This paper adopts a sample partitioning strategy grounded in labeled pixels. Specifically, all valid labeled pixels are first identified using the ground-truth label maps provided for each dataset; subsequently, these pixels are partitioned into training, validation, and test sets according to a predefined sampling ratio and class-balanced sample counts. Specifically, for the Indian Pines dataset, 10% of the samples are allocated to the training set, another 10% to the validation set, and the remaining samples constitute the test set; the corresponding class-wise sample distribution is presented in Table 1. For both the Pavia University and Xuzhou datasets, 0.5% of the samples are assigned to the training set and 0.5% to the validation set, with the remainder reserved for testing; their respective class distributions are detailed in Tables 2 and 3. In the case of the WHU-Hi-LongKou dataset, 0.2% of the samples serve as the training set, 0.2% as the validation set, and the rest as the test set; the associated class distribution is summarized in Table 4. Additionally, for each sample, a 15×15 spatial image patch centered on the corresponding pixel is extracted as input to the spatial branch, and the sample label is determined by the class of the central pixel.

To ensure the fairness of comparative experiments, all experiments in this paper were conducted on an Ubuntu operating system. The hardware platform comprised an NVIDIA GeForce RTX 4090 GPU designed in the United States, and an Intel Core i7-13700KF CPU designed by Intel Corporation(USA). The proposed method was implemented in Python (version 3.9.19), with models constructed using the PyTorch framework (version 2.3.0) and code developed and debugged in PyCharm (version 2024.1). The Adam optimizer was employed for training the proposed method, with a fixed learning rate of 0.0005. For baseline methods, optimizers and learning rates were configured strictly according to the optimal hyperparameters reported in their original publications. In all experiments, the batch size was set to 64, the number of training epochs to 100, and the number of retained spectral bands to 30. Classification performance was evaluated using three quantitative metrics: overall accuracy (OA), average accuracy (AA), and kappa coefficient (k). Higher values of all three metrics indicate superior classification performance. To mitigate the influence of stochastic variation, each experiment was independently repeated five times;

final results are reported as the mean across these repetitions, with the best-performing result in each comparison highlighted in bold.

Table 1. Sample distribution of Indian Pines dataset.

No.	Train	Val	Test
C01	5	4	37
C02	142	143	1143
C03	83	83	664
C04	24	23	190
C05	48	49	386
C06	73	73	584
C07	3	3	22
C08	48	48	382
C09	2	2	16
C10	97	97	778
C11	245	246	1964
C12	59	59	475
C13	21	20	164
C14	126	127	1012
C15	38	39	309
C16	10	9	74
Total	1024	1025	8200

Table 2. Sample distribution of Pavia University dataset.

No.	Train	Val	Test
C01	33	33	6565
C02	93	93	18,463
C03	10	11	2078
C04	15	16	3033
C05	7	6	1332
C06	25	25	4979
C07	7	6	1317
C08	18	19	3645
C09	5	5	937
Total	213	214	42,349

Table 3. Sample distribution of Xuzhou dataset.

No.	Train	Val	Test
C01	132	131	26,133
C02	20	20	3987
C03	14	14	2755
C04	26	26	5162
C05	66	66	13,052
C06	12	12	2412
C07	35	35	6920
C08	24	24	4729
C09	15	16	3039
Total	344	344	68,189

Table 4. Sample distribution of WHU-Hi-LongKou dataset.

No.	Train	Val	Test
C01	69	69	34,373
C02	16	17	8341
C03	6	6	3019
C04	127	126	62,959
C05	9	8	4134
C06	23	24	11,807
C07	134	134	66,788
C08	14	15	7095
C09	11	10	5208
Total	409	409	203,724

3.3. Comparative Experiment

To comprehensively evaluate the performance of the proposed method, this paper selects nine state-of-the-art classification approaches for comparative analysis: 3D-CNN [53], DFFN [54], RSSAN [26], SPRN [55], SF [39], SSFTT [43], GAHT [44], morph [45], and GSC-ViT [47]. These baseline methods encompass diverse architectural paradigms—including 3D convolutional neural networks and attention-based models—thereby enabling multi-faceted comparisons and robust verification of the efficacy and generalizability of the proposed approach.

(1) Classification Maps and Experimental Results for the Indian Pines Dataset

Table 5 presents the quantitative evaluation metrics of the proposed method and comparative methods on the Indian Pines dataset. Figure 8 displays the classification maps generated by all methods on the Indian Pines dataset. As demonstrated by the experimental results, the proposed method achieves superior classification performance, with overall accuracy (OA), average accuracy (AA), and kappa coefficient reaching 99.04%, 98.79%, and 98.90%, respectively.

Table 5. Classification results of different methods on the Indian Pines dataset.

Class	3D-CNN	DFFN	RSSAN	SPRN	SF	SSFTT	GAHT	Morph	GSC-ViT	Ours
C01	93.51	87.03	34.59	95.14	92.97	92.97	87.57	91.35	91.35	**98.92**
C02	92.41	95.00	90.66	**98.60**	96.10	98.36	98.29	97.90	97.53	98.41
C03	95.93	96.93	85.63	98.83	98.19	98.01	98.61	97.83	98.61	**99.13**
C04	93.47	95.89	64.32	97.58	97.58	97.16	**99.16**	97.47	97.58	**99.16**
C05	96.58	94.72	90.31	97.10	95.60	96.37	97.36	**97.82**	96.17	97.36
C06	99.21	97.47	92.60	99.01	**99.52**	99.28	98.87	98.12	98.70	99.38
C07	98.18	75.45	36.36	**100**	94.55	95.45	85.45	99.09	**100**	95.45
C08	99.79	99.74	97.23	99.63	98.74	99.84	99.90	**100**	99.58	**100**
C09	85.00	65.00	18.75	**100**	95.00	55.00	22.50	77.50	93.75	98.75
C10	91.95	93.73	85.66	96.25	94.16	96.12	95.66	94.86	94.11	**99.10**
C11	98.15	97.95	96.44	98.25	97.87	98.72	97.93	97.85	98.59	**99.02**
C12	96.63	93.47	80.42	96.46	96.42	97.09	97.68	96.67	97.98	**98.78**
C13	**99.88**	97.44	90.12	96.71	99.76	98.29	97.07	98.17	96.95	**99.88**
C14	99.33	99.70	96.86	99.92	99.21	99.70	99.82	99.74	99.80	**99.98**
C15	97.09	94.76	82.27	98.32	**99.29**	98.32	97.73	96.57	96.25	98.45
C16	94.05	90.00	44.59	89.19	95.14	90.54	85.41	89.73	82.70	**98.92**
OA(%)	96.52	96.49	89.87	98.19	97.45	98.12	97.83	97.67	97.75	**99.04**
AA(%)	95.70	92.14	74.18	97.56	96.88	94.45	91.19	95.67	96.23	**98.79**
$k \times 100$	96.03	95.99	88.43	97.94	97.10	97.86	97.53	97.34	97.43	**98.90**

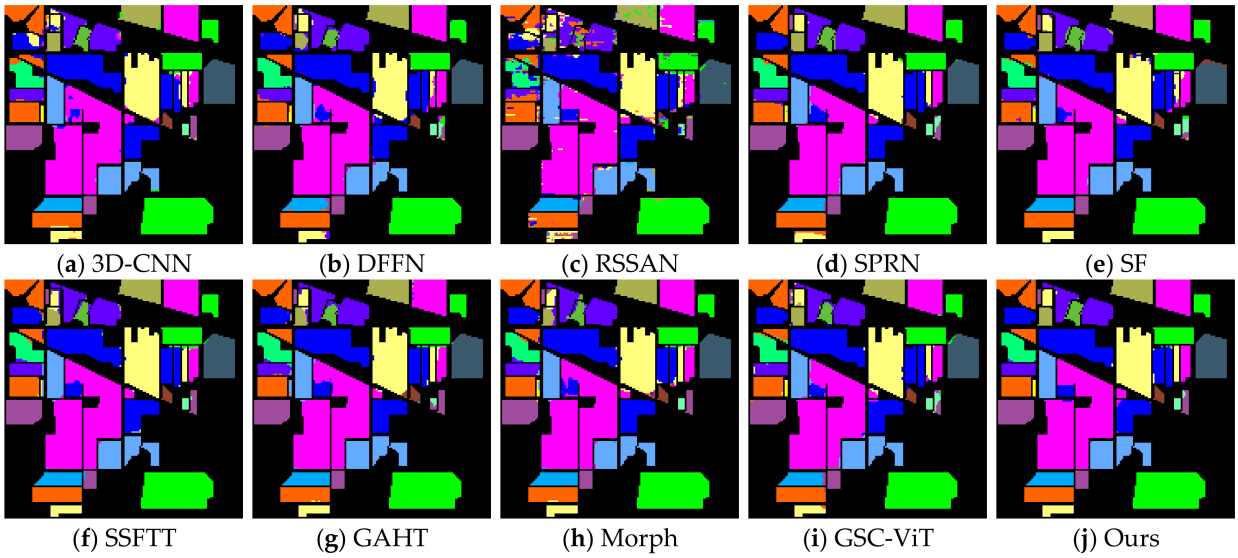


Figure 8. Classification maps of all methods on the Indian Pines dataset.

SSFTT, a transformer-based model, achieves a high classification accuracy while maintaining lightweight architecture, attaining an overall accuracy (OA) of 98.12%. Notably, SPRN—a CNN-based method—achieves the second-best classification performance among all compared approaches, with an OA of 98.19%. In terms of average accuracy (AA), the proposed method outperforms SPRN, SF, and GSC-ViT by 1.23%, 1.91%, and 2.56%, respectively; the improvement over other baseline methods is even more pronounced. Visual inspection of the classification maps reveals that RSSAN produces maps heavily contaminated with noise. In contrast, the classification maps generated by the remaining comparative methods exhibit comparable visual quality. The proposed method yields classification maps that are particularly clear and well structured in target regions of Classes 1, 9, and 16 and demonstrates superior preservation of spatial structural consistency in complex land-cover boundary areas.

(2) Classification Maps and Experimental Results for the Pavia University Dataset

Table 6 and Figure 9 present the classification accuracy and classification maps, respectively, of the proposed method and all comparative methods on the Pavia University dataset. As demonstrated by the experimental results, the proposed method achieves superior classification performance, with overall accuracy (OA), average accuracy (AA), and kappa coefficient reaching 98.19%, 96.92%, and 97.61%, respectively—outperforming all baseline methods.

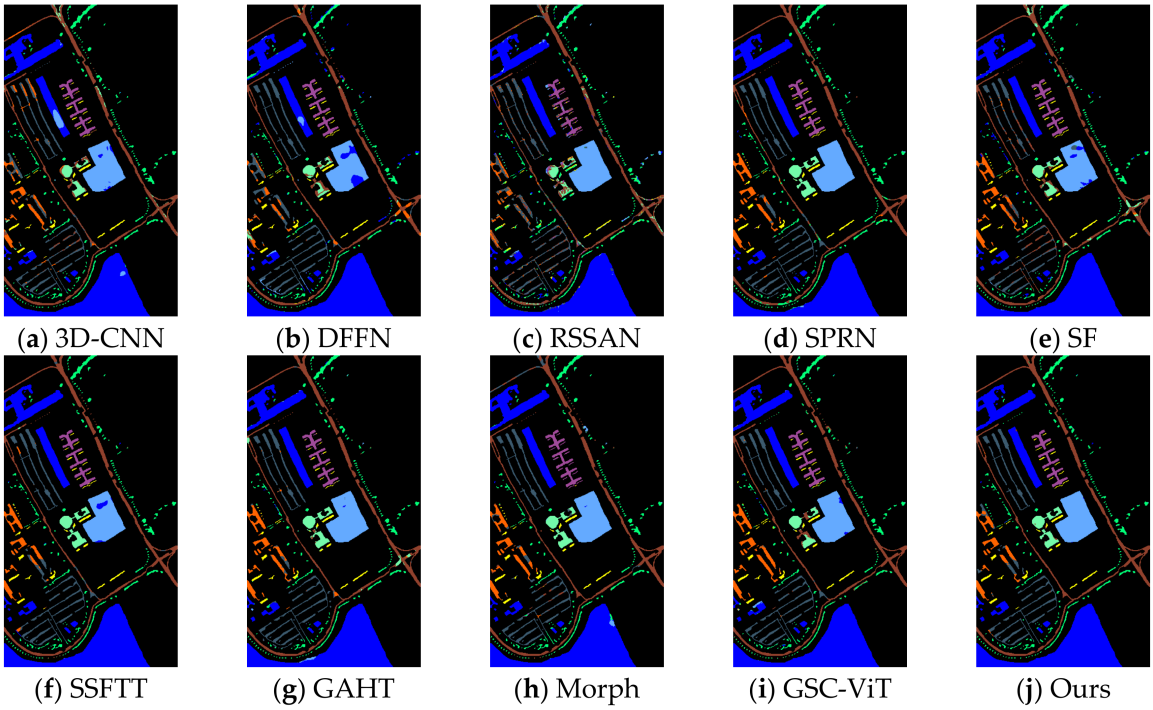
The second-highest classification accuracy is achieved by SSFTT, with overall accuracy (OA), average accuracy (AA), and kappa coefficient reaching 97.51%, 95.45%, and 96.69%, respectively. Regarding per-class accuracy, the proposed method also demonstrates superior performance: the classification accuracy for the vast majority of classes exceeds 98%, particularly for classes C05, C07, and C09—where it significantly outperforms all comparative methods.

As revealed in the classification maps, the proposed method exhibits no conspicuous misclassification in Class 6, whereas most baseline methods show varying degrees of classification errors for this class. For Classes 7 and 9, GAHT and GSC-ViT suffer from evident class confusion; in contrast, the proposed method yields spatially contiguous classification regions with well-defined boundaries.

(3) Classification Maps and Experimental Results for the Xuzhou Dataset

Table 6. Classification results of different methods on the Pavia University dataset.

Class	3D-CNN	DFFN	RSSAN	SPRN	SF	SSFTT	GAHT	Morph	GSC-ViT	Ours
C01	91.37	88.41	86.14	96.43	89.50	96.53	96.09	95.02	97.22	**98.44**
C02	98.78	98.55	99.10	99.80	99.30	**99.94**	99.59	99.55	99.41	99.50
C03	61.05	47.16	38.40	86.57	75.92	88.90	85.14	87.42	**92.44**	88.73
C04	95.74	81.85	78.40	85.87	87.97	92.29	88.07	88.50	87.46	**96.95**
C05	99.94	99.13	82.54	99.43	99.92	99.82	**99.95**	99.29	98.63	99.71
C06	93.47	74.47	95.81	99.22	93.81	98.12	98.28	**99.55**	97.10	99.17
C07	66.48	57.89	26.77	98.95	77.62	**99.79**	98.47	95.85	87.06	97.28
C08	88.92	92.17	73.54	96.59	83.18	96.30	94.59	92.15	**97.81**	96.63
C09	85.51	75.62	46.70	83.61	84.97	87.32	70.72	80.94	91.29	**95.84**
OA(%)	92.83	88.13	86.11	96.89	92.82	97.51	96.27	96.29	96.88	**98.19**
AA(%)	86.81	79.47	69.71	94.05	88.02	95.45	92.32	93.14	94.27	**96.92**
$k \times 100$	90.46	83.96	81.40	95.87	90.44	96.69	95.05	95.08	95.85	**97.61**



(**a**) 3D-CNN (**b**) DFFN (**c**) RSSAN (**d**) SPRN (**e**) SF

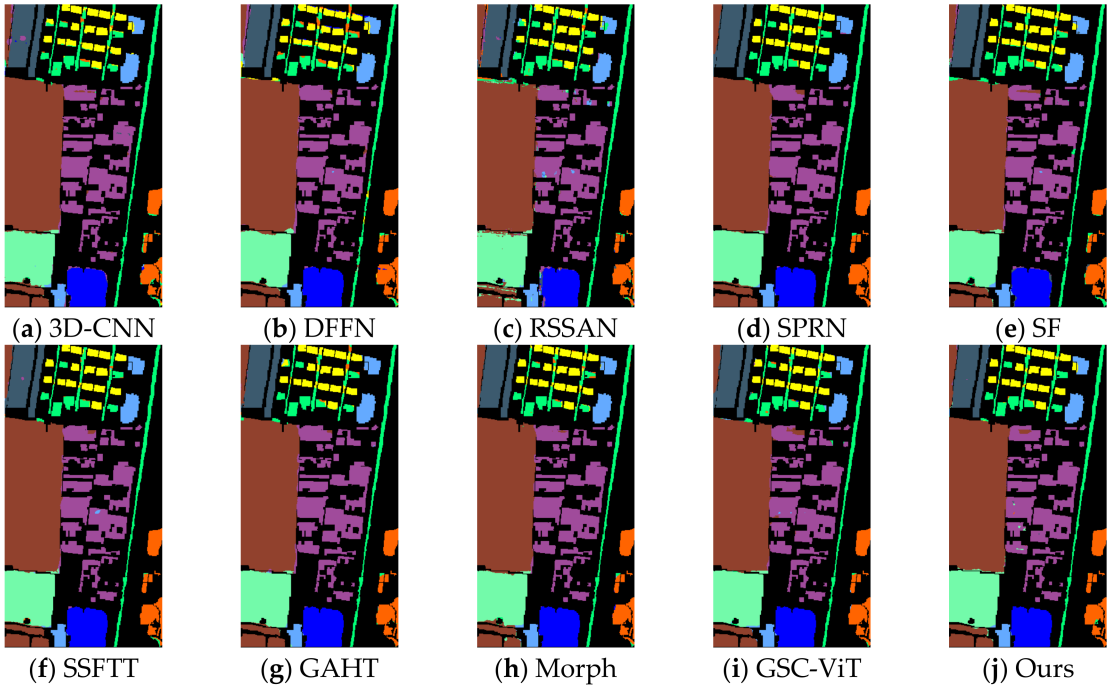
(**f**) SSFTT (**g**) GAHT (**h**) Morph (**i**) GSC-ViT (**j**) Ours

Figure 9. Classification maps of all methods on the Pavia University dataset.

The classification accuracy and corresponding classification maps of the proposed method and all comparative methods on the Xuzhou dataset are presented in Table 7 and Figure 10, respectively. The proposed method achieves overall accuracy (OA), average accuracy (AA), and kappa coefficient values of 98.60%, 98.60%, and 98.22%, respectively—outperforming all competing approaches. It attains the highest per-class accuracy for the majority of categories (C01, C02, C03, C04, C07, and C09). SSFTT ranks second, with OA, AA, and kappa values of 97.91%, 97.99%, and 97.36%, respectively. Overall, CNN-based methods—such as DFFN and RSSAN—exhibit comparatively inferior performance, achieving OAs of only 95.45% and 94.80%. Notably, 3D-CNN demonstrates classification performance nearly comparable to that of SF.

Table 7. Classification results of different methods on the Xuzhou dataset.

Class	3D-CNN	DFFN	RSSAN	SPRN	SF	SSFTT	GAHT	Morph	GSC-ViT	Ours
C01	97.32	96.41	96.21	97.00	97.03	97.31	97.03	97.49	97.69	**98.49**
C02	96.88	99.07	92.99	98.39	95.79	99.29	99.01	98.89	98.99	**99.59**
C03	91.41	88.57	90.30	98.26	93.16	97.66	98.35	96.35	97.40	**99.25**
C04	93.21	85.61	88.80	95.37	91.64	95.17	92.25	91.24	93.58	**96.67**
C05	**99.17**	97.50	97.45	98.94	98.78	99.05	99.08	99.11	97.71	98.69
C06	90.90	91.04	88.49	97.44	98.14	98.18	95.27	**98.26**	97.72	97.66
C07	97.21	96.74	95.75	98.71	98.24	98.41	98.73	96.88	98.58	**99.40**
C08	97.13	98.85	95.64	**99.96**	98.90	99.72	99.19	99.90	99.42	99.31
C09	94.69	91.77	89.58	94.65	96.36	97.10	93.72	93.52	94.70	**98.33**
OA(%)	96.73	95.45	94.80	97.67	96.99	97.91	97.34	97.32	97.52	**98.60**
AA(%)	95.32	93.95	92.80	97.64	96.45	97.99	96.96	96.85	97.31	**98.60**
$k \times 100$	95.85	94.22	93.41	97.05	96.19	97.36	96.63	96.60	96.86	**98.22**



(**a**) 3D-CNN (**b**) DFFN (**c**) RSSAN (**d**) SPRN (**e**) SF

(**f**) SSFTT (**g**) GAHT (**h**) Morph (**i**) GSC-ViT (**j**) Ours

Figure 10. Classification maps of all methods on the Xuzhou dataset.

Furthermore, visual inspection of the classification maps reveals that RSSAN's output contains conspicuous noise artifacts. While SF produces clean and well-defined target-class regions, isolated misclassified pixels (i.e., speckle noise) remain present. In contrast, the classification map generated by the proposed method exhibits superior spatial consistency and significantly reduced noise—even within fine-scale land-cover patches, such as Class 4 and Class 9, where accurate delineation is particularly challenging.

(4) Classification Maps and Experimental Results for the WHU-Hi-LongKou Dataset

The classification accuracy and corresponding classification maps of the proposed method and all comparative methods on the WHU-Hi-LongKou dataset are presented in Table 8 and Figure 11, respectively. The proposed method achieves the best overall classification performance on this dataset, with overall accuracy (OA), average accuracy (AA), and kappa coefficient reaching 99.14%, 97.72%, and 98.87%, respectively.

Table 8. Classification results of different methods on the WHU-Hi-LongKou dataset.

Class	3D-CNN	DFFN	RSSAN	SPRN	SF	SSFTT	GAHT	Morph	GSC-ViT	Ours
C01	99.58	98.90	98.68	99.61	99.75	**99.85**	**99.85**	99.65	99.82	99.71
C02	92.11	92.93	93.63	97.83	94.90	**98.70**	95.15	97.12	96.95	95.43
C03	82.57	46.37	73.31	89.62	92.67	93.28	91.66	88.14	93.70	**95.93**
C04	98.94	97.16	97.59	99.16	98.37	99.18	98.85	**99.23**	**99.23**	99.09
C05	74.27	25.53	63.64	**95.87**	83.60	94.92	86.08	86.62	89.87	95.09
C06	99.38	97.32	97.04	99.00	99.41	99.34	98.96	98.67	99.33	**99.73**
C07	**99.97**	99.60	99.38	99.61	**99.97**	99.67	99.54	99.79	99.78	**99.97**
C08	92.53	89.25	78.20	87.97	88.73	89.00	86.85	85.99	89.81	**97.61**
C09	94.30	79.64	76.13	81.33	85.93	82.52	86.86	82.22	87.49	**96.91**
OA(%)	98.05	95.16	95.90	98.27	98.01	98.49	98.01	98.05	98.52	**99.14**
AA(%)	92.63	80.74	86.40	94.45	93.70	95.16	93.76	93.05	95.11	**97.72**
$k \times 100$	97.43	93.62	94.60	97.72	97.38	98.02	97.38	97.43	98.06	**98.87**

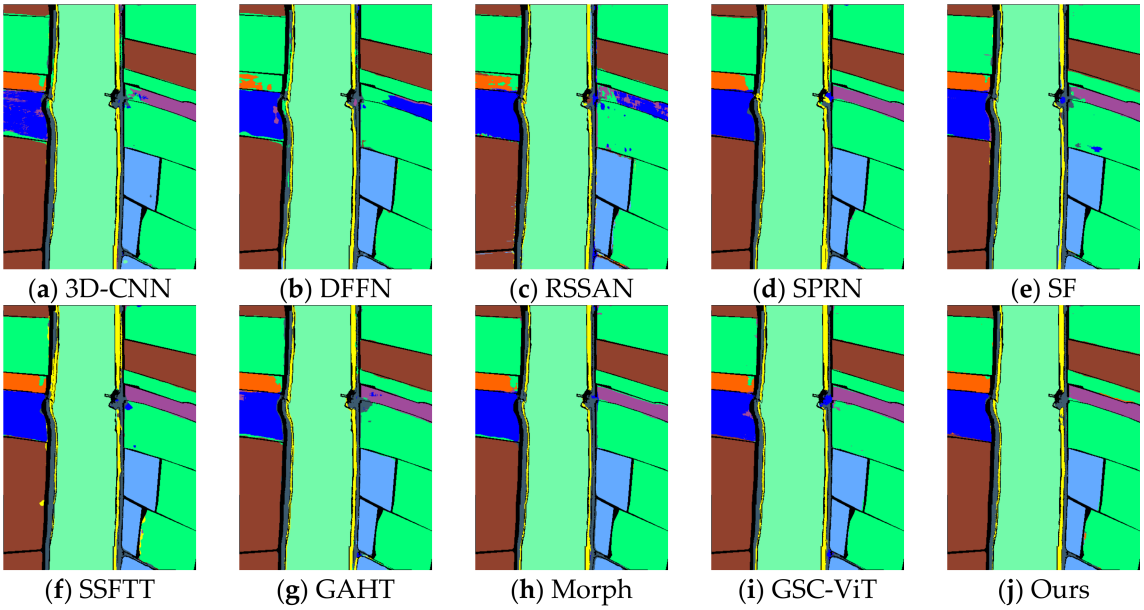


Figure 11. Classification maps of all methods on the WHU-Hi-LongKou dataset.

The second- and third-best performing methods are GSC-ViT and SSFTT. Specifically, GSC-ViT attains an OA of 98.52%, an AA of 95.11%, and a kappa coefficient of 98.06%. A per-class accuracy analysis reveals that the proposed method significantly outperforms all comparative methods on classes C03, C05, C08, and C09. Notably, on class C09, the proposed method achieves classification accuracy improvements of 9.42% and 14.39% over GSC-ViT and SSFTT, respectively.

In the classification maps, RSSAN exhibits the highest level of noise. Although SPRN and morph yield classification maps with less noise than RSSAN, spurious color patches persist near class boundaries. Both GSC-ViT and SSFTT produce misclassifications—specifically, pixels belonging to Class 3 are erroneously assigned to Class 4. In contrast, the classification map generated by the proposed method in this paper demonstrates significantly reduced intra-class noise. No prominent inter-class misclassifications are observed; moreover, land-cover boundaries are sharp, well defined, and spatially continuous—particularly for target classes such as Class 8 and Class 9, where classification performance is notably superior.

3.4. Parameter Experiment

(1) Patch Size

To investigate the impact of varying image patch sizes on classification performance, experiments were conducted across four datasets using different patch sizes while keeping all other hyperparameters fixed. The results are presented in Figure 12a. As the patch size increases, the overall accuracy of the proposed method first rises and then declines. This trend arises because excessively small patches provide insufficient spatial contextual information, thereby limiting the model's ability to learn discriminative features and degrading classification performance. Conversely, overly large patches introduce substantial redundancy and noise, which hinder effective feature learning. Comprehensive analysis of the experimental results indicates that a patch size of 15 yields optimal classification accuracy across all four datasets. Consequently, this value is adopted as the default patch size in all subsequent experiments.

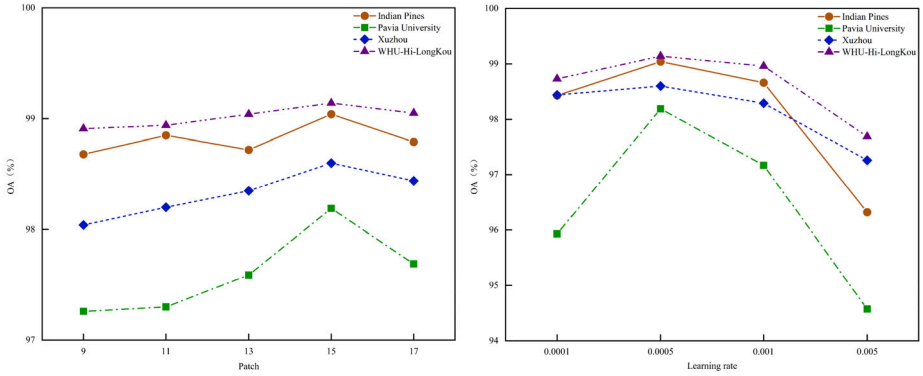


Figure 12. Parameter experiments on four datasets: (**a**) patch size; (**b**) learning rate.

(2) Learning Rate

The learning rate is a critical hyperparameter in the model optimization process, and its value directly influences both the convergence speed and generalization capability of the model. To investigate this effect, we conducted experiments on four benchmark datasets using four distinct learning rates while keeping all other hyperparameters fixed; the results are presented in Figure 12b. An excessively small learning rate leads to minimal parameter updates per iteration, resulting in slow convergence and an increased risk of getting trapped in local optima. Conversely, an overly large learning rate causes excessively aggressive parameter updates, destabilizing the training process and preventing proper convergence. Based on a comprehensive analysis of the experimental results, a learning rate of 0.0005 consistently yields the best classification performance across all four datasets. Consequently, this value is adopted as the default learning rate for all subsequent experiments.

(3) Number of Groups

To investigate the impact of the number of groups in grouped deformable convolution and grouped convolution on the model's classification performance, experiments were conducted on four datasets with all other parameters held constant while varying the number of groups. In the experimental setup, the number of groups was set identically for both grouped deformable convolution and grouped convolution. The overall accuracy of the proposed method under different group numbers is presented in Table 9. On the Indian Pines, Xuzhou, and WHU-Hi-LongKou datasets, the model achieves its highest overall classification accuracy when the number of groups is set to eight. In contrast, on the Pavia

University dataset, the best overall classification accuracy is attained when the number of groups is two.

Table 9. Overall accuracies (%) of the proposed method on the four datasets with different group numbers.

Dataset	Number of Groups				
	2	**4**	**8**	**16**	**32**
Indian Pines	98.89	98.82	**99.04**	98.84	98.51
Pavia University	**98.19**	97.86	97.91	97.39	97.52
Xuzhou	98.40	98.44	**98.60**	98.54	98.58
WHU-Hi-LongKou	99.04	98.96	**99.14**	99.00	99.04

3.5. Ablation Study

To further assess the contribution of individual components to the model's classification performance, ablation studies were conducted on the Pavia University and Xuzhou datasets. These experiments systematically evaluated the roles of three key modules: the deformable spatial feature fusion module (DSFFM), the global spatial semantic aggregation module (GSSAM), and the spectral transformer module (SpeFormer). The corresponding results are summarized in Table 10.

Table 10. Ablation study of the proposed method on the Pavia University and Xuzhou datasets.

Scheme	Module			Pavia University			Xuzhou		
	DSFFM	**GSSAM**	**SpeFormer**	**OA(%)**	**AA(%)**	$k \times 100$	**OA(%)**	**AA(%)**	$k \times 100$
1	√	√	√	**98.19**	**96.92**	**97.61**	**98.60**	**98.60**	**98.22**
2	GC	√	√	97.60	95.92	96.82	97.75	97.27	97.14
3	√	-	√	96.99	95.10	95.99	97.67	97.27	97.05
4	√	√	-	96.35	92.81	95.15	97.58	97.40	96.93

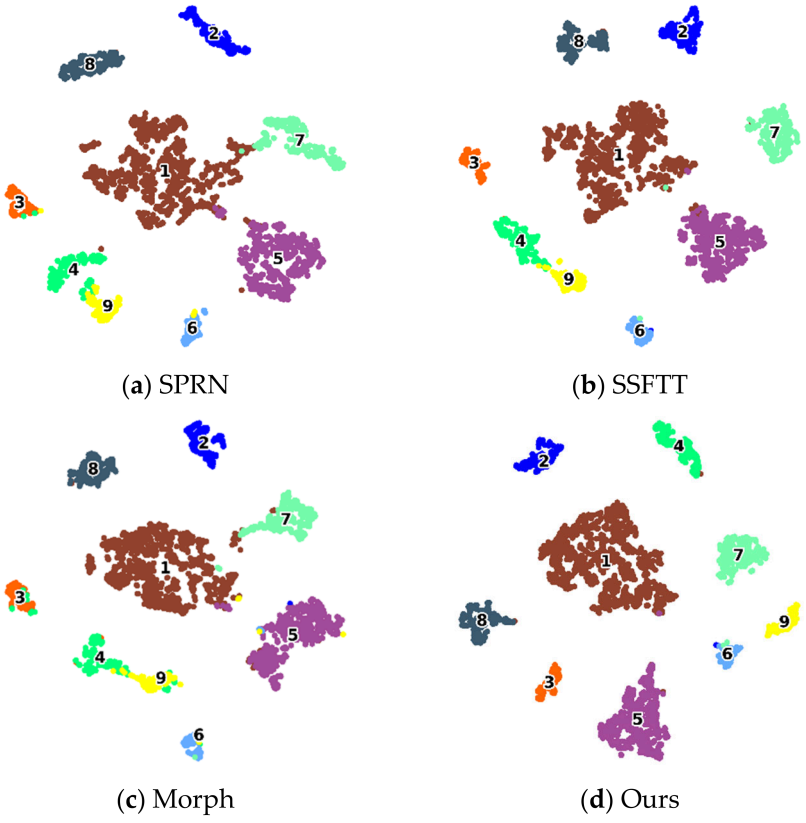
Specifically, Scheme 1 represents the complete network architecture integrating DSFFM, GSSAM, and SpeFormer, achieving the best classification performance across both datasets. Scheme 2 replaces the grouped deformable convolution (GDC) in DSFFM with a standard grouped convolution (GC), while keeping all other modules unchanged, to analyze the impact of GDC on classification performance. As shown in the results, Scheme 2 yields consistently lower OA, AA, and kappa values on both datasets: its OA drops to 97.6% and 97.75% on Pavia University and Xuzhou, respectively—representing reductions of 0.59% and 0.85% compared to Scheme 1. This demonstrates that fine-grained spatial structure modeling via deformable convolution significantly enhances classification accuracy. Scheme 3 removes GSSAM while retaining DSFFM and SpeFormer to evaluate the role of the second-stage spatial semantic aggregation. Under this configuration, OA declines to 96.99% and 97.67%, AA to 95.1% and 97.27%, and kappa to 95.99% and 97.05% on the Pavia University and Xuzhou datasets, respectively. These results underscore the importance of strengthening semantic representation capability and global spatial relationship modeling through GSSAM. Scheme 4 eliminates the spectral branch and uses only the spatial branch for classification, aiming to analyze the contribution of spectral features to the joint spatial–spectral representation learning. Its OA falls to 96.35% and 97.58% on the two datasets—decreasing by 1.84% and 1.02% relative to Scheme 1. This confirms that joint representation learning of spectral and spatial information contributes meaningfully to improved classification performance.

In summary, GDC, GSSAM, and SpeFormer positively contribute to classification performance from the perspectives of spatial structure enhancement, global spatial semantic

aggregation, and spectral feature learning, respectively. Ablation results further validate the effectiveness of each model.

3.6. t-SNE Visualization Analysis

This paper employs t-distributed stochastic neighbor embedding (t-SNE) for two-dimensional visualization of high-dimensional features extracted by different methods on the Xuzhou dataset, thereby further analyzing the discriminative capability of the proposed approach. The experimental results are presented in Figure 13. t-SNE preserves the local structure of samples while mapping high-dimensional features into a low-dimensional space, enabling intuitive visualization of the distribution of different classes in the feature space. Compared with the t-SNE visualizations generated by SPRN, SSFTT, and morph, the proposed method exhibits significantly reduced feature overlap among Class 4, Class 9, Class 1, and Class 7. Moreover, intra-class sample distributions become more compact, and inter-class boundaries become markedly sharper. These improvements stem from the proposed method's dual advantages—fine-grained modeling of local spatial details via grouped deformable convolution, and joint modeling of spatial and spectral features—both of which collectively enhance the discriminability of learned feature representations.



(**a**) SPRN (**b**) SSFTT

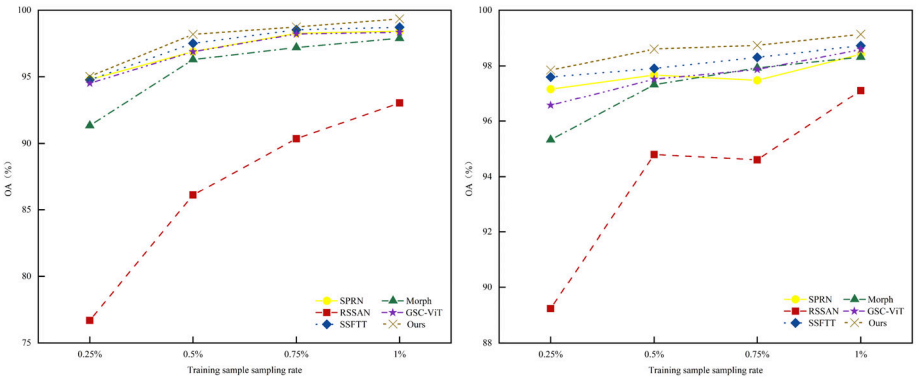
(**c**) Morph (**d**) Ours

Figure 13. t-SNE visualization results of features on the Xuzhou dataset.

3.7. The Impact of Training Sample Ratios

To evaluate the performance of the proposed method under varying training sample sizes and to investigate the impact of training set size on classification accuracy, this paper conducts experiments with different training sample sampling rates. Specifically, on the Pavia University and Xuzhou datasets, training sets are constructed by randomly sampling 0.25%, 0.5%, 0.75%, and 1% of the labeled samples, respectively, while the remaining samples are used for testing. Comparative results are presented in Figure 14. As shown, the proposed method consistently achieves superior classification performance across all

sampling ratios and outperforms all baseline methods. Moreover, although the classification accuracy of all methods improves with increasing training sample size, the proposed method maintains significantly higher and more stable accuracy throughout. These findings further substantiate the effectiveness and robustness of the proposed approach.



(**a**) Pavia University (**b**) Xuzhou

Figure 14. Impact of different training sample ratios on experimental results on two datasets.

3.8. Efficiency Analysis

To further evaluate the computational efficiency of the proposed method, this paper conducts a comparative analysis from four perspectives: training time, inference time, number of model parameters, and floating point operations (FLOPs). Experiments are performed on four hyperspectral datasets—Indian Pines, Pavia University, Xuzhou, and WHU-Hi-LongKou—and the corresponding quantitative results are summarized in Table 11.

Table 11. Computational efficiency of all methods on the four datasets.

Methods	Indian Pines		Pavia University		Xuzhou		WHU-Hi-LongKou		Params (M)	FLOPs (M)
	Train (s)	Test (s)	Train (s)	Test (s)	Train (s)	Test (s)	Train (s)	Test (s)		
3D-CNN	**43.62**	**0.90**	**9.14**	**6.59**	**14.72**	**4.49**	**16.42**	**8.07**	0.30	45.09
DFFN	67.93	2.37	20.49	28.66	27.01	16.68	34.21	28.72	0.40	89.65
RSSAN	57.28	2.02	15.99	21.79	21.47	12.20	25.82	20.67	**0.05**	10.41
SPRN	45.84	1.32	10.83	15.44	16.34	8.59	20.35	16.58	0.18	39.52
SF	119.25	6.66	29.67	69.99	41.51	45.68	48.80	71.34	0.14	**4.44**
SSFTT	50.43	1.63	13.61	16.42	19.08	11.51	23.30	18.52	0.15	16.91
GAHT	69.79	2.46	19.29	25.98	30.10	15.61	35.58	31.52	4.44	164.97
Morph	101.22	4.24	23.52	46.91	35.72	28.64	41.55	47.82	0.06	8.64
GSC-ViT	89.15	3.48	21.17	34.02	33.55	21.46	38.26	33.40	0.22	37.15
Ours	103.44	4.89	21.63	36.54	39.15	26.25	45.74	41.49	0.72	138.40

Integrating classification accuracy with computational cost enables a more holistic evaluation of the proposed method. Taking the Pavia University dataset as an example, our method achieves an overall accuracy of 98.19%, representing a 0.68-percentage-point improvement over SSFTT's OA of 97.51%. However, this gain is accompanied by increased computational overhead: the number of parameters rises from 0.15 M to 0.72 M, FLOPs increase from 16.91 M to 138.40 M, and inference time extends from 16.42 s to 36.54 s. These results indicate that, relative to this lightweight baseline, the accuracy improvement entails a non-negligible increase in computational demand.

In contrast, compared to GAHT, our method achieves a 1.92-percentage-point OA improvement (from 96.27% to 98.19%) while simultaneously reducing the parameter count from 4.44 M to 0.72 M and lowering FLOPs from 164.97 M to 138.40 M. Nevertheless, inference time increases from 25.98 s to 36.54 s. This observation underscores the fact that reductions in parameter count and FLOPs do not invariably translate into shorter actual runtime—execution efficiency remains highly dependent on implementation-specific factors such as memory access patterns, kernel optimization, and hardware utilization.

Overall, the proposed method attains high classification accuracy under the current experimental settings; however, its computational cost exceeds that of certain lightweight alternatives. In practical deployment, model selection should carefully balance accuracy requirements, available computational resources, and real-time inference constraints. Further optimization to reduce computational overhead thus remains a key direction for future work.

4. Discussion

The experimental results presented above demonstrate that the proposed method delivers substantial classification performance improvements across all four benchmark datasets. As summarized in Tables 5–8, the method achieves an overall accuracy (OA) of 99.04% on the Indian Pines dataset—surpassing the second-best performer, SPRN (OA = 98.19%), by 0.85 percentage points. On the Pavia University dataset, it attains an OA of 98.19%, exceeding the second-highest result achieved by SSFTT (OA = 97.51%) by 0.68 percentage points. Similarly, the proposed method yields the highest classification accuracy on both the Xuzhou and WHU-Hi-LongKou datasets. Visual inspection of the classification maps across all four datasets further corroborates its superiority over competing approaches. Specifically, the maps generated by our method exhibit sharply delineated boundaries between land-cover classes and significantly reduced misclassification—particularly among spectrally or spatially similar features—compared to alternative methods.

Overall, under the experimental settings adopted in this study, transformer-based classification methods achieve consistently strong performance across most benchmark datasets and demonstrate clear overall advantages. This superiority stems primarily from the self-attention mechanism inherent to transformers, which enables dynamic modeling of interdependencies among spatially distant positions and thereby effectively captures long-range contextual relationships—significantly enhancing the model's capacity to represent global spatial structure. For hyperspectral image classification, accurate land-cover identification depends not only on local spatial structure and neighborhood context but also on comprehensive exploitation of discriminative spectral characteristics across the full band continuum. Spatial information encodes geometric properties—including shape, texture, and spatial distribution—whereas spectral information delivers distinctive, material-specific response signatures across wavelengths. Motivated by these complementary yet essential modalities, this paper proposes a dual-branch spatial–spectral joint representation learning framework designed to explicitly and independently model spatial and spectral features.

In spatial branches, conventional convolutional operations employ fixed sampling positions—a design that proves inadequate for modeling objects in hyperspectral images, which often exhibit irregular geometries and complex spatial distributions. To address this limitation, this paper introduces grouped deformable convolution in the first stage. This mechanism learns spatially adaptive sampling offsets to dynamically adjust convolutional receptive fields, thereby significantly enhancing the model's capacity to capture intricate local spatial structures. Furthermore, a transformer module is integrated to explicitly model long-range contextual dependencies across diverse spatial locations. Building upon the

enhanced spatial representations generated in the first stage, the second stage focuses on high-level feature transformation and global semantic aggregation—rather than redundant refinement of sampling positions. Accordingly, standard grouped convolution replaces grouped deformable convolution; this enables efficient intra-group feature transformation while eliminating the computational overhead associated with offset prediction. Subsequently, a classification token–based transformer architecture is employed to holistically integrate information from heterogeneous spatial regions, yielding discriminative, globally coherent spatial–semantic representations. Through this hierarchical design, the learning process transitions seamlessly—from fine-grained spatial structure enhancement in the first stage to comprehensive global spatial–semantic aggregation in the second.

Meanwhile, recognizing that hyperspectral data exhibit both global dependencies across distant spectral bands and local correlations among adjacent bands, this paper integrates a transformer architecture with a convolutional feed-forward network within the spectral branch. Specifically, the transformer captures long-range spectral dependencies through self-attention mechanisms, while the convolutional feed-forward network strengthens local feature interactions among neighboring bands via spectral convolutions.

To further clarify the distinctions between the proposed method and existing approaches, we analyze three key aspects: deformable spatial sampling, CNN–transformer hybrid architectures, and dual-branch spatial–spectral modeling. Regarding deformable convolution-based hyperspectral classification models, the superpixel guided deformable convolution network leverages superpixel segmentation to inform the sampling locations of deformable convolutions, thereby enabling adaptive spatial sampling aligned with object boundaries. In contrast, our DSFFM partitions feature maps along the channel dimension into multiple disjoint groups and assigns group-specific, learnable spatial offsets—allowing it to capture heterogeneous spatial structures across different feature subspaces. The resulting enhanced features are subsequently fed into the GSSAM for holistic spatial semantic integration. Therefore, the main distinctions between our approach and these methods lie in the grouping sampling mechanism and the integration strategy for subsequent spatial relationship modeling. For the CNN–transformer hybrid architecture, SSFTT first extracts shallow spatial–spectral features using both 3D and 2D convolutions and then applies Gaussian-weighted tokenization followed by a transformer module to learn discriminative feature representations. This paper likewise integrates convolutional operations with self-attention mechanisms; however, it introduces two sequential, functionally distinct stages within the spatial branch: the DSFFM captures irregular local structures and their intrinsic spatial relationships, while the GSSAM further refines and fuses spatial contextual information. For the dual-branch hyperspectral transformer, CASST fosters synergistic interaction between spatial and spectral features through cross-attention mechanisms and weight sharing. This work adopts two independent yet complementary branches—spatial and spectral—with a primary design objective of enhancing spatial feature representation. Specifically, the spatial branch jointly models irregular spatial structures and performs subsequent spatial semantic aggregation. In parallel, the spectral branch employs self-attention to capture long-range dependencies along the spectral sequence of the central pixel, while further strengthening local feature interactions via a convolutional feed-forward network. Finally, the outputs from both branches are concatenated for downstream classification. Consequently, the core innovations of this method reside in its two-stage spatial feature enhancement within the spatial branch and its integrated modeling of local–global spectral features.

In the subsequent t-SNE visualization results (Figure 13), the proposed method likewise exhibits superior performance. This improvement can be attributed to the dual-branch architecture: the spatial branch enhances spatial structural integrity and aggregates global

semantic information, thereby fully exploiting structural and contextual cues across diverse spatial locations; meanwhile, the spectral branch complements this by capturing inter-band spectral correlations. The synergistic fusion of these complementary spatial and spectral features significantly improves the discriminability of distinct land-cover classes within the feature space.

Furthermore, given the high cost associated with acquiring labeled samples in hyperspectral image classification, this paper systematically investigates the influence of varying training sample ratios on model classification performance. As illustrated in Figure 14, on both the Pavia University and Xuzhou datasets, the overall classification accuracy of all evaluated methods generally improves as the training sample ratio increases from 0.25% to 1%. In comparison with other state-of-the-art approaches, our method consistently achieves superior classification performance across all tested training sample ratios—particularly demonstrating notable advantages under low-labeling regimes. This robustness is likely attributable to the dual-branch architecture's complementary integration of spatial and spectral information. Collectively, these experimental results further substantiate the strong generalization capability and label efficiency of our method under data-scarce conditions.

5. Conclusions

This paper proposes a dual-branch spatial–spectral transformer with a spatial feature enhancement model, within a unified spatial–spectral joint representation learning framework. In the spatial branch, a deformable spatial feature fusion module is introduced to effectively model local details embedded in complex spatial structures, thereby enhancing the model's capacity to represent spatial characteristics of irregular ground objects. Subsequently, a global spatial semantic aggregation module performs higher-level semantic integration over the spatial features generated in the preceding stage, yielding more discriminative spatial representations. In the spectral branch, a transformer architecture is employed to model spectral sequences, augmented by a convolutional feed-forward network to further strengthen the discriminability of spectral features. Finally, spatial and spectral features are fused to achieve joint representation learning that synergistically leverages both spatial and spectral information. Experimental results demonstrate that the proposed method achieves superior classification performance across multiple benchmark hyperspectral datasets. However, this study adopts random pixel-level partitioning within the same scene, which may introduce spatial overlap between training and test image patches. Consequently, the reported results primarily reflect classification performance under this specific partitioning scheme, and the model's generalization capability on spatially independent regions remains to be rigorously validated. Moreover, current evaluations rely predominantly on global classification metrics and qualitative classification maps, with insufficient quantitative assessment of model performance in land-cover boundary areas. To address these limitations, our future work will enhance experimental evaluation in two key directions: first, by implementing spatially disjoint—e.g., block-based or grid-aligned—training, validation, and test partitions, incorporating adequate spatial separation to ensure no patch-level overlap across subsets, thereby enabling a more robust assessment of spatial generalization; second, by incorporating boundary-specific metrics—including boundary accuracy and boundary F-score—to quantitatively evaluate the method's effectiveness in land-cover transition zones.

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

Funding: This work was supported by the National Natural Science Foundation of China under Grant No. 62671109 and by the Joint-Plan Fund of Liaoning Provincial Science and Technology Program under Grant No. 2024011291-JH2/1026.

Data Availability Statement: All datasets used in the article are publicly available: https://ieee-dataport.org/documents/hyperspectral-remote-sensing-datasets-indian-pines-pavia-university-botswana-and-salinas (accessed on 6 February 2026), https://ieee-dataport.org/documents/xuzhou-hyspex-dataset (accessed on 6 February 2026), and http://rsidea.whu.edu.cn/resource_WHUHi_sharing.htm (accessed on 6 February 2026).

Acknowledgments: The authors also thank Youdao Translate (accessed on 27 August 2026) for its linguistic translation and polishing assistance during the preparation of this manuscript. The authors have fully reviewed and revised the generated content and assume full responsibility for all content of this publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gevaert, C.M.; Suomalainen, J.; Tang, J.; Kooistra, L. Generation of Spectral-Temporal Response Surfaces by Combining Multispectral Satellite and Hyperspectral UAV Imagery for Precision Agriculture Applications. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3140–3146. [CrossRef]
2. Zhang, L.; Zhang, M.; Huang, J.; Zhang, C.; Ye, F.; Pan, W. A New Approach for Mineral Mapping Using Drill-Core Hyperspectral Image. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5511705. [CrossRef]
3. Yuan, J.; Wang, S.; Wu, C.; Xu, Y. Fine-Grained Classification of Urban Functional Zones and Landscape Pattern Analysis Using Hyperspectral Satellite Imagery: A Case Study of Wuhan. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3972–3991. [CrossRef]
4. Zhang, M.; Li, W.; Zhao, X.; Liu, H.; Tao, R.; Du, Q. Morphological Transformation and Spatial-Logical Aggregation for Tree Species Classification Using Hyperspectral Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5501212. [CrossRef]
5. Xia, X.; Liu, W.; Wang, L.; Sun, J. HSIFoodIngr-64: A Dataset for Hyperspectral Food-Related Studies and a Benchmark Method on Food Ingredient Retrieval. *IEEE Access* **2023**, *11*, 13152–13162. [CrossRef]
6. Blanzieri, E.; Melgani, F. Nearest Neighbor Classification of Remote Sensing Images with the Maximal Margin Principle. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1804–1811. [CrossRef]
7. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
8. Ham, J.; Chen, Y.; Crawford, M.M.; Ghosh, J. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 492–501. [CrossRef]
9. Tong, F.; Zhang, Y. Spectral-Spatial and Cascaded Multilayer Random Forests for Tree Species Classification in Airborne Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5613311. [CrossRef]
10. Licciardi, G.; Marpu, P.R.; Chanussot, J.; Benediktsson, J.A. Linear Versus Nonlinear PCA for the Classification of Hyperspectral Data Based on the Extended Morphological Profiles. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 447–451. [CrossRef]
11. Villa, A.; Benediktsson, J.A.; Chanussot, J.; Jutten, C. Hyperspectral Image Classification with Independent Component Discriminant Analysis. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4865–4876. [CrossRef]
12. Du, Q. Modified Fisher's Linear Discriminant Analysis for Hyperspectral Imagery. *IEEE Geosci. Remote Sens. Lett.* **2007**, *4*, 503–507. [CrossRef]
13. Ye, Z.; Bai, L.; Tan, L. Hyperspectral image classification based on Gabor features and decision fusion. In Proceedings of the 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, China, 2–4 June 2017; pp. 478–482.
14. Zhang, C.; Zhu, S.; Xue, D.; Sun, S. Gabor Filter-Based Multi-Scale Dense Network Hyperspectral Remote Sensing Image Classification Technique. *IEEE Access* **2023**, *11*, 114146–114154. [CrossRef]
15. Benediktsson, J.A.; Palmason, J.A.; Sveinsson, J.R. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 480–491. [CrossRef]
16. Chen, Y.; Nasrabadi, N.M.; Tran, T.D. Hyperspectral Image Classification Using Dictionary-Based Sparse Representation. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3973–3985. [CrossRef]
17. Zhao, X.; Wang, L.; Zhang, Y.; Han, X.; Deveci, M.; Parmar, M. A review of convolutional neural networks in computer vision. *Artif. Intell. Rev.* **2024**, *57*, 99. [CrossRef]
18. Bai, J.; Huang, S.; Xiao, Z.; Li, X.; Zhu, Y.; Regan, A.C.; Jiao, L. Few-Shot Hyperspectral Image Classification Based on Adaptive Subspaces and Feature Transformation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5523917. [CrossRef]

19. Feng, J.; Wu, X.; Shang, R.; Sui, C.; Li, J.; Jiao, L.; Zhang, X. Attention Multibranch Convolutional Neural Network for Hyperspectral Image Classification Based on Adaptive Region Search. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5054–5070. [CrossRef]
20. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*, 258619. [CrossRef]
21. Yang, J.; Zhao, Y.-Q.; Chan, J.C.-W. Learning and Transferring Deep Joint Spectral-Spatial Features for Hyperspectral Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4729–4742. [CrossRef]
22. 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]
23. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
24. Roy, S.K.; Manna, S.; Song, T.; Bruzzone, L. Attention-Based Adaptive Spectral-Spatial Kernel ResNet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7831–7843. [CrossRef]
25. Cai, Y.; Chang, Z.; Wang, L.; Wang, X.; Liu, W.; Zhuo, Y. DSCD: A Dual-Stream Cross-Domain Network for Few-Shot Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 11234–11246. [CrossRef]
26. 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]
27. Paoletti, M.E.; Moreno-Álvarez, S.; Xue, Y.; Haut, J.M.; Plaza, A. AAtt-CNN: Automatic Attention-Based Convolutional Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5511118. [CrossRef]
28. Liu, Q.; Dong, Y.; Zhang, Y.; Luo, H. A Fast Dynamic Graph Convolutional Network and CNN Parallel Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5530215. [CrossRef]
29. Zhao, C.; Zhu, W.; Feng, S. Superpixel Guided Deformable Convolution Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2022**, *31*, 3838–3851. [CrossRef] [PubMed]
30. Zhuang, J.; Wang, K.; Yuan, Z.; Yan, Y. Frequency domain iterative clustering for boundary-preserving superpixel segmentation. In *Applied Soft Computing*; Elsevier: Amsterdam, The Netherlands, 2026; p. 114717.
31. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
32. Paoletti, M.E.; Moreno-Álvarez, S.; Haut, J.M. Multiple Attention-Guided Capsule Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5520420. [CrossRef]
33. Feng, J.; Yu, H.; Wang, L.; Cao, X.; Zhang, X.; Jiao, L. Classification of Hyperspectral Images Based on Multiclass Spatial-Spectral Generative Adversarial Networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5329–5343. [CrossRef]
34. Wan, S.; Gong, C.; Zhong, P.; Pan, S.; Li, G.; Yang, J. Hyperspectral Image Classification with Context-Aware Dynamic Graph Convolutional Network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 597–612. [CrossRef]
35. 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*, 6000–6010.
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 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
37. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
38. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
39. 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*, 5518615. [CrossRef]
40. Liao, W.; Wang, F.; Zhao, H. Hyperspectral Image Classification Using Attention-Only Spatial-Spectral Network Based on Transformer. *IEEE Access* **2024**, *12*, 93677–93688. [CrossRef]
41. 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]
42. He, X.; Chen, Y.; Lin, Z. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]
43. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral-Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5522214. [CrossRef]
44. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral Image Classification Using Group-Aware Hierarchical Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5539014. [CrossRef]
45. Roy, S.K.; Deria, A.; Shah, C.; Haut, J.M.; Du, Q.; Plaza, A. Spectral-Spatial Morphological Attention Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5503615. [CrossRef]

46. Zhang, J.; Meng, Z.; Zhao, F.; Liu, H.; Chang, Z. Convolution Transformer Mixer for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6014205. [CrossRef]
47. 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*, 5511817. [CrossRef]
48. Zhao, J.; Wang, J.; Ruan, C.; Dong, Y.; Huang, L. Dual-Branch Spectral-Spatial Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–18. [CrossRef]
49. Arshad, T.; Zhang, J. A Light-Weighted Spectral-Spatial Transformer Model for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 12008–12019. [CrossRef]
50. Zhang, Y.; Wang, Z.; Huang, M.; Li, M.; Zhang, J.; Wang, S.; Zhang, J.; Zhang, H. S2DBFT: Spectral-Spatial Dual-Branch Fusion Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5525517. [CrossRef]
51. Wang, Y.; Tai, M.; Liu, Z.; Shu, Z.; Xiong, F.; Yu, Z. Mid-Range Convolutional Modulated Transformer Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5532315. [CrossRef]
52. Zhong, Z.; Liang, C.; Yang, M.; Wang, D. LCTNet: Lightweight Convolution-Transformer Network for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 7844–7857. [CrossRef]
53. Ben Hamida, A.; Benoit, A.; Lambert, P.; Ben Amar, C. 3-D Deep Learning Approach for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4420–4434. [CrossRef]
54. 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]
55. Zhang, X.; Shang, S.; Tang, X.; Feng, J.; Jiao, L. Spectral Partitioning Residual Network with Spatial Attention Mechanism for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5507714. [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.