

Article

Multi-Scale Lightweight Spectral Attention Network for Hyperspectral Image Classification

Yanjuan Wang ^{1,2} , Yanxue Zhang ¹, Yi Liu ¹, Yujia Hou ¹, Feifan Jiang ¹, Ziyang Zhao ¹, Mingyang Shang ¹, Wanming Zhao ^{2,3} and Fengqi Li ^{1,*} 

¹ School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian 116028, China; wangyanjuan@djtu.edu.cn (Y.W.); 15095039753@163.com (Y.Z.); ly9404121107@163.com (Y.L.); noneye@163.com (Y.H.); angyang0827@163.com (F.J.); z13832715152@163.com (Z.Z.); aurora13912@163.com (M.S.)

² State Key Laboratory of Maritime Technology and Safety, Dalian Maritime University, Dalian 116026, China; zhaowm@dlmu.edu.cn

³ College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

* Correspondence: lifengqi@dlut.edu.cn

Highlights

What are the main findings?

- A lightweight MLSANet framework is proposed for hyperspectral image classification, reducing model complexity while maintaining strong spectral–spatial feature representation.
- The LPPA and PADA modules are developed to enhance discriminative spectral responses and capture pixel-wise multi-scale spatial contexts, respectively.

What are the implications of the main findings?

- The lightweight design of MLSANet provides an effective strategy for improving the accuracy–efficiency trade-off in hyperspectral image classification.
- The results provide a useful reference for designing efficient hyperspectral classification models by coordinating redundancy reduction, attention-based feature enhancement, and structure-aware representation.

Abstract

Hyperspectral remote sensing technology provides a unique data foundation for accurate land-cover identification by acquiring continuous spectral information of land cover. To address the issues of large parameter count, high computational complexity, and insufficient utilization of spectral–spatial features in existing hyperspectral image classification methods, a Multi-Scale Lightweight Spectral Attention Network (MLSANet) is proposed. The proposed network consists of three components: a lightweight feature extraction layer, an efficient attention enhancement layer, and a structure-aware representation layer. The lightweight feature extraction layer reduces computational overhead through principal component analysis (PCA) for dimensionality reduction, spatial downsampling, depth-wise separable convolution, and low-rank factorized convolution. The efficient attention enhancement layer comprises two novel modules: the Lightweight Pruning Parameterized Attention (LPPA) module and the Pixel-wise Adaptive Dilated Attention (PADA) module. The LPPA module prunes features based on an energy function and introduces a learnable scaling factor to adaptively enhance salient features. The PADA module utilizes pixel-wise attention and multi-scale dilated convolution to finely fuse spectral–spatial features. Furthermore, the structure-aware representation layer integrates a graph convolutional network with a lightweight multi-layer perceptron to model global structural relationships, thereby further improving feature discriminability. Experimental results on three public



Academic Editor: Qiang Li

Received: 26 May 2026

Revised: 19 July 2026

Accepted: 22 July 2026

Published: 24 July 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

datasets, including Indian Pines, Pavia Centre, and Salinas, demonstrate that the proposed MLSANet model achieves competitive classification accuracy compared with recent representative methods, attaining overall accuracies of 99.01%, 98.66%, and 95.88%, respectively, while simultaneously exhibiting lower parameter counts and computational costs. This confirms the favorable trade-off between accuracy and efficiency achieved by our model.

Keywords: hyperspectral; image classification; lightweight; attention mechanism; spectral–spatial features

1. Introduction

Hyperspectral remote sensing technology enables the acquisition of continuous spectral information of land cover across dozens to hundreds of narrow bands ranging from the visible to shortwave infrared wavelengths, using imaging spectrometers mounted on satellites or airborne platforms [1]. Compared with traditional multispectral remote sensing, hyperspectral remote sensing achieves the integration of spectral and spatial information, which means that each pixel contains not only the spatial location of the land cover but also a continuous spectral curve that allows fine characterization of the spectral reflectance properties of land covers [2]. Hyperspectral data exhibit three main characteristics: first, high spectral resolution down to the nanometer scale, enabling the detection of subtle spectral features of land covers; second, a large number of continuous bands form a three-dimensional data cube; and third, a large data volume with high inter-band correlation, resulting in substantial redundant information [3]. Based on these unique advantages, hyperspectral technology has been widely applied in fields such as precision agriculture, environmental monitoring, mineral exploration, and defense security [4–6].

Hyperspectral image classification, as a core task in the interpretation of hyperspectral remote sensing data, aims to assign an accurate land cover label to each pixel, and it serves as a key step in converting massive spectral data into usable information [7]. The research significance of this task can be understood from the following three perspectives. In agriculture, high-precision classification enables fine identification of crop types and early detection of pests and diseases, providing decision support for precision agriculture [8]. In the environmental domain, real-time monitoring and classification of water bodies, soil, and other environmental elements facilitates the timely detection of pollution sources and the assessment of ecological risks [9]. In defense, hyperspectral classification technology can effectively identify camouflaged targets and military facilities, thereby enhancing battlefield situational awareness [10]. However, the high dimensionality of hyperspectral data, the high correlation between bands, and the scarcity of labeled samples pose substantial challenges to classification tasks, making the design of efficient and accurate hyperspectral image classification methods a current research hotspot and difficulty [11].

In recent years, deep learning techniques have substantially advanced the development of hyperspectral image classification [12]. Convolutional neural networks (CNNs), owing to their advantages in local feature extraction and weight sharing, are capable of effectively capturing the local spatial structures and spectral correlations in hyperspectral images and thus have been widely adopted for joint spectral–spatial feature modeling [13]. Researchers have subsequently proposed various CNN-based hyperspectral classification models, such as one-dimensional CNNs for extracting spectral sequence features [14], two-dimensional CNNs for extracting spatial features [15], and three-dimensional CNNs for simultaneously extracting both spectral and spatial features [16]. Building upon these efforts, the HybridSN model was proposed to combine 3D and 2D convolutions, first

employing 3D convolutions to extract joint spectral–spatial features and then using 2D convolutions for deep spatial modeling, thereby further improving feature fusion capability while maintaining computational efficiency [17]. However, the receptive field of CNNs is limited by the kernel size, making it difficult to capture global contextual information; moreover, as network depth increases, computational complexity rises significantly, and overfitting becomes more likely [18]. To address this issue, researchers have introduced recurrent neural networks (RNNs), leveraging their sequence modeling capability to capture the continuity of the spectral dimension, for example, by using long short-term memory (LSTM) networks or gated recurrent units (GRUs) to model long-range dependencies in spectral curves [19]. Nevertheless, RNNs are difficult to train, the problems of vanishing and exploding gradients persist, and their computational efficiency remains low.

Since its introduction by Vaswani et al. in 2017, the Transformer architecture has achieved breakthroughs in natural language processing and computer vision [20]. Its core self-attention mechanism is capable of effectively modeling long-range dependencies in sequences, overcoming the limitation of the local receptive field inherent to CNNs. Compared with RNNs, Transformers significantly improve training efficiency through parallel computation while avoiding the vanishing gradient problem. In addition, the multi-head attention mechanism enables the model to learn richer contextual information from different feature subspaces. These advantages make the Transformer naturally suitable for handling spectral sequences and spatial relationships in hyperspectral data, offering a new paradigm for hyperspectral image classification [21]. Among the proposed approaches, the Spectral–Spatial Feature Tokenization Transformer (SSFTT) proposed by Sun et al. is a representative hybrid architecture model [22]. This model first employs 3D and 2D convolutions to extract shallow spectral–spatial features, transforming the three-dimensional hyperspectral data into two-dimensional feature maps; then, it converts the feature maps into sequences of semantic tokens through feature tokenization; finally, it uses a Transformer encoder to perform high-level semantic modeling on the token sequences, learning global contextual dependencies. By organically combining CNNs and Transformers, SSFTT effectively integrates the strengths of local feature extraction and global context modeling, achieving excellent performance in hyperspectral image classification. However, this model still suffers from the following limitations: first, its relatively large parameter count and high computational complexity restrict its deployment and application in resource-constrained scenarios; second, the utilization of spectral information remains insufficient, failing to fully utilize the strong correlations between bands and the subtle variations in spectral curves, thereby limiting the model’s feature representation capability.

With the rapid development of state space models (SSMs), Mamba, as a representative SSM architecture, has increasingly been introduced into hyperspectral image classification. Unlike Transformers, which rely on self-attention mechanisms for global modeling, Mamba enables efficient learning of long-range dependencies through selective state space modeling while maintaining relatively low computational complexity [23]. To accommodate the spectral–spatial characteristics of hyperspectral data, a variety of Mamba-based hyperspectral classification methods have been developed. For instance, SS-Mamba was among the first to explore the application of state space models to spectral–spatial feature modeling in hyperspectral images, where spectral and spatial sequences were constructed to capture long-range dependencies [24]. Subsequently, S^2 Mamba [25] further integrated dual-branch spatial–spectral state space modeling with a feature fusion mechanism, thereby enhancing the model’s capability to represent complex spectral–spatial information. Hierarchical Spatial–Spectral Mamba adopts hierarchical spectral–spatial modeling to improve the adaptability of Mamba to hyperspectral classification tasks [26]. In addition, WD-SSMamba incorporated a wavelet decomposition strategy to strengthen

the integration of frequency-domain information and state space modeling, leading to improved classification performance [27]. Nevertheless, these methods generally rely on relatively complex sequence modeling architectures and may still incur additional computational overhead while achieving strong representation capabilities. In recent years, foundation models such as SpectralGPT have also opened up new directions for hyperspectral representation learning [28]. However, they typically require large-scale pretraining datasets and substantial computational resources, which limits their practical applicability in resource-constrained scenarios.

To address the aforementioned issues, we propose a Multi-Scale Lightweight Spectral Attention Network (MLSANet), inspired by the spectral–spatial feature modeling strategy of SSFTT. The proposed network comprises three main components: a lightweight feature extraction layer, an efficient attention enhancement layer, and a structure-aware representation layer. The lightweight feature extraction layer is specifically designed to overcome the challenges inherent in initial spectral–spatial feature extraction for hyperspectral image classification, including high computational overhead, excessive local redundant responses, and the susceptibility of discriminative features to background noise interference. This layer first applies principal component analysis (PCA) to the original hyperspectral data for spectral dimensionality reduction, which preserves the dominant spectral information while reducing the computational cost caused by redundant bands and suppressing noise interference associated with low-variance components. Subsequently, spatial downsampling is employed to reduce the spatial size of the input features, thereby decreasing the computational burden of subsequent convolution operations at an early stage. Furthermore, depthwise separable convolution is adopted to achieve lightweight spectral–spatial feature extraction by decoupling spatial feature extraction from channel-wise feature fusion. Meanwhile, asymmetric convolution factorization is introduced to decompose a standard two-dimensional convolution into consecutive one-dimensional directional convolutions, further reducing redundant computation in spatial convolution. Through these designs, the lightweight feature extraction layer not only reduces parameter counts and computational complexity but also yields a more compact basic spectral–spatial feature representation, while simultaneously attenuating the interference of irrelevant background noise on the subsequent attention enhancement process. Secondly, the efficient attention enhancement layer consists of two modules: the Lightweight Pruning Parameterized Attention (LPPA) and the Pixel-wise Adaptive Dilated Attention (PADA), which are designed to enhance feature representation from the spectral and spatial dimensions, respectively. Specifically, LPPA is designed to address the strong inter-band correlation and substantial redundant spectral responses in hyperspectral images. Its core idea is to estimate the importance of different spectral responses through channel-wise energy modeling. In particular, an energy score is constructed according to the deviation of each channel feature from the mean response, thereby converting channel response discrepancies into measurable importance weights. Based on this mechanism, LPPA further suppresses low-contribution redundant responses via threshold-based pruning and performs lightweight recalibration on the retained responses using learnable scaling factors, enabling the model to focus more on discriminative spectral channels. PADA, in turn, addresses insufficient spatial context modeling from the spatial perspective by leveraging dilated convolutions to expand the local receptive field and capture richer neighboring contextual information. Multi-scale spatial feature extraction is adopted to capture local structural variations at different scales, while a pixel-wise attention gating mechanism adaptively modulates responses at each spatial location, enabling the model to focus more on class-discriminative key spatial regions while suppressing the influence of weakly relevant and background areas. Finally, the structure-aware representation layer further models spatial relationships among the

enhanced features. Considering that convolution and attention operations still have limited capability in representing latent structural correlations among feature nodes, a graph convolutional network (GCN) is introduced to capture spatial dependencies beyond local neighborhoods. On this basis, a lightweight multilayer perceptron (LMLP) is constructed to perform compact mapping and category discrimination on the structure-aware features, thereby improving the discriminability of the final feature representation while maintaining a controlled parameter scale.

To validate the effectiveness of the proposed method, experiments were conducted on three public datasets, namely Indian Pines, Pavia Centre, and Salinas. The experimental results demonstrate that the proposed MLSANet significantly reduces the number of model parameters and computational complexity while achieving competitive classification accuracy compared with recent representative methods. This validates its favorable balance between accuracy and efficiency, offering an effective solution for hyperspectral image classification in resource-constrained environments.

The main contributions of this paper are as follows:

1. A Multi-Scale Lightweight Spectral Attention Network (MLSANet) is proposed. Through the synergistic design of a lightweight feature extraction layer, an efficient attention enhancement layer, and a structure-aware representation layer, the network significantly reduces computational complexity while improving spectral-spatial feature discriminability. Experimental results demonstrate that MLSANet enhances spectral-spatial feature representation while maintaining a low number of model parameters and computational complexity.
2. A Lightweight Pruning Parameterized Attention (LPPA) module is proposed to suppress redundant responses and select discriminative channels along the spectral dimension. Unlike conventional channel attention mechanisms that mainly rely on global pooling or channel-weight recalibration, LPPA evaluates the importance of spectral responses from the perspective of channel-wise energy modeling. Specifically, it converts the deviation of channel features from the mean response into measurable energy scores. Based on this energy estimation, LPPA further introduces a threshold-based pruning mechanism to suppress low-contribution spectral responses and employs lightweight learnable scaling factors to adaptively recalibrate the retained responses. In this way, LPPA more effectively alleviates the strong inter-band correlation and response redundancy commonly observed in hyperspectral data.
3. A Pixel-wise Adaptive Dilated Attention (PADA) module is proposed to enhance spatial contextual representation and perform pixel-level response modulation along the spatial dimension. PADA enlarges the spatial receptive field through convolutions with different dilation rates, captures local spatial structures at different scales and directions using multi-scale directional convolutions, and adaptively modulates the fused spatial responses through pixel-wise attention gating. This design enhances the contextual representation capability of critical spatial regions with low computational overhead, thereby compensating for the insufficient spatial modeling capability caused by relying solely on local convolutions or channel attention.

2. Related Work

2.1. Lightweight Networks and Multi-Scale Feature Fusion

In hyperspectral image classification, lightweight model design and multi-scale feature fusion are seen as two key ways to improve efficiency and accuracy. In recent years, a series of lightweight network architectures have been proposed to handle the high-dimensional redundancy in hyperspectral data. Depthwise separable convolution is an effective lightweight technique. It has been widely adopted to reduce model parameters

and computational cost. For example, in the MobileNet family, computational cost is reduced by using depthwise separable convolutions. As a result, a good balance between accuracy and efficiency is achieved in image classification tasks. In the field of hyperspectral image classification, the HybridSN model proposed by Roy et al. combines 3D and 2D convolutions to extract multi-scale features, which improves classification accuracy but still has a relatively large number of parameters [17]. To further reduce model size, the SSRN proposed by Zhong et al. introduces residual connections and batch normalization, which alleviates the vanishing gradient problem and reduces the number of parameters [18]. However, the aforementioned methods still have limitations in multi-scale feature modeling, which makes it difficult to capture both fine details of targets at different scales and global contextual information.

To address multi-scale feature fusion in hyperspectral image classification, inspiration has been drawn from feature pyramid networks, and cross-layer feature fusion structures have been designed for hyperspectral images. FNet was proposed by Hu et al., in which a feature interaction module is used to enhance the fusion of target-level and pixel-level features, thus improving multi-scale feature representation. FBRNet was developed by Fu et al. It adopts an anchor-free design and adds a multi-scale feature enhancement strategy to make multi-scale recognition more robust. In hyperspectral image classification, a Progressive Multi-scale Multi-attention Fusion Network (PMMF) was proposed by Tang et al. This network repeatedly extracts features from different branches using an “extract-stack-extract-fusion” strategy. It also designs a multi-attention fusion module, which applies the most suitable attention mechanism based on the feature characteristics of each branch. As a result, multi-scale features are sufficiently fused, and feature map information is enriched. CLFR-Det was proposed by Liu et al. It deeply fuses features of different scales through a cross-layer feature refinement mechanism, which improves the representation of fine spatial structures. Nevertheless, existing methods still have two problems during feature fusion: insufficient semantic consistency and high computational cost. Therefore, more efficient lightweight multi-scale feature fusion strategies are needed.

2.2. Attention Mechanisms and Spectral–Spatial Feature Enhancement

Attention mechanisms dynamically adjust feature weights to emphasize informative regions while suppressing irrelevant background responses, thereby improving classification performance. In hyperspectral image classification, due to the complex spectral characteristics of land cover and their strong correlations, attention mechanisms are widely used to improve the discriminability of spectral–spatial features. Traditional channel attention and spatial attention are important methods for feature enhancement in hyperspectral image classification. SENet introduces channel attention, which adaptively recalibrates channel features through global average pooling and fully connected layers, enabling the model to focus on important spectral channels. CBAM further combines channel and spatial attention, where attention weights are computed along the two dimensions, respectively, thereby enhancing both channel selection and spatial region focusing capabilities [29]. To reduce the large parameter count in traditional attention modules, ECA-Net replaces fully connected layers with one-dimensional convolution, achieving efficient channel attention with almost no increase in parameters [30]. In hyperspectral image classification, a lightweight RetinaNet was proposed by Miao et al., which simplifies the structure of attention modules to improve classification performance without increasing computational cost [31]. A multi-level feature attention mechanism was designed by Yang et al., which assigns weights to the spectral and spatial dimensions separately, allowing the model to adaptively focus on important feature channels and spatial regions [32]. A multi-scale self-distillation network was developed by Liu et al., which combines self-distillation tech-

niques with lightweight design, guides feature learning through attention mechanisms, and enhances the recognition capability for complex classes [33]. A shallow feature reconstruction module was proposed by Bai et al., which employs spatial attention to enhance semantic information and addresses classification challenges in complex backgrounds [34].

In recent years, the self-attention mechanism of the Transformer architecture has attracted much attention due to its powerful global modeling capability. Unlike traditional channel or spatial attention, self-attention can model long-range dependencies between any two elements in a sequence. The original Transformer architecture proposed by Vaswani et al. uses a multi-head self-attention mechanism, which allows the model to learn rich contextual information from different feature subspaces [20]. In computer vision, the Vision Transformer (ViT) proposed by Dosovitskiy et al. processes images by dividing them into sequences of patches, showing the potential of Transformers for image classification tasks [35]. Inspired by this, researchers have started to apply Transformers to hyperspectral classification. HSI-BERT was proposed by He et al., which introduces multi-head self-attention into hyperspectral pixel sequence modeling. It treats the spectral vector of each pixel as a sequence input, thereby capturing long-range spectral dependencies effectively [36]. Considering the strong continuity along the spectral dimension, SpectralFormer was developed by Hong et al. It learns local sequences of adjacent spectral bands, which enhances the local continuity representation of spectral features while keeping the structural advantages of Transformers [37]. To deal with the high computational cost of Transformers, GAHT was proposed by Mei et al. It restricts multi-head self-attention to a local spectral-spatial context and reduces computational cost through a grouping strategy, while maintaining good feature representation capability [38]. Furthermore, a spectral-spatial transformer network was proposed by Yang et al., which models spectral and spatial information using a Transformer-based framework [39]. A Spectral-Spatial Transformer Network (SSTN) was proposed by Zhong et al., which combines spectral correlation with spatial attention modules and replaces traditional convolution operations [40].

Spectral-spatial feature enhancement is important for improving hyperspectral image classification. Various approaches have been explored to enhance these features from different perspectives. Hang et al. proposed Attention-Aided CNNs, which enhance the discriminative spectral channels and spatial location responses through spectral attention and spatial attention subnetworks, respectively, thereby improving the joint spectral-spatial feature representation capability [41]. Huang et al. introduced a dual-branch attention-assisted CNN, which enhances feature representation in hyperspectral image classification through a spectral-spatial branch combined with an attention mechanism, while also alleviating the issues of large parameter counts and information redundancy in deep CNNs [42]. Liang et al. developed HSI-Mixer, which extracts spectral-spatial features using convolution and further models feature interactions via a Mixer representation [43]. Subsequently, Transformer and attention mechanisms have been further introduced into the task of spectral-spatial feature enhancement. Li and Ye proposed STNet, which enhances the fusion of spatial context and fine-grained spectral information through spectral-spatial attention decoupling and an adaptive gating mechanism [44]. Ahmad et al. presented a hyperspectral classification method that combines multi-scale Transformer with spatial attention, enhancing spectral-spatial feature representation through multi-scale feature extraction and spatial attention [45]. In addition, Mahdavi-pour et al. proposed TGDHTL, which integrates Transformer, graph convolutional networks, and diffusion enhancement mechanisms to improve global relationship modeling and transferability in hyperspectral classification [46]. Nevertheless, while enhancing spectral-spatial feature representation, existing methods still face challenges such as spectral redundancy, insufficient spatial context modeling, or high model complexity.

3. Methodology

To address the issues of large parameter count, high computational complexity, and insufficient utilization of spectral–spatial features in existing hyperspectral image classification methods, we propose a Multi-Scale Lightweight Spectral Attention Network (MLSANet). The overall framework of the proposed network is illustrated in Figure 1. The model mainly consists of three components. The first is a lightweight feature extraction layer, in which spatial downsampling is applied to significantly reduce the spatial dimensions of the input features, depthwise separable convolution is used to replace standard 3D convolution, and a low-rank convolution kernel decomposition structure (e.g., the combination of 1×3 and 3×1 convolutions) is introduced in the 2D convolution stage, reducing parameter count and computational overhead while preserving feature extraction capability. The second component is an efficient attention enhancement layer, which comprises the proposed Lightweight Pruning Parameterized Attention (LPPA) module and the Pixel-wise Adaptive Dilated Attention (PADA) module. The LPPA module suppresses redundant information through pruning and lightweight parameterization while retaining global saliency modeling capability, whereas the PADA module employs pixel-wise attention and dilated convolution to replace conventional large-kernel convolution operations, achieving efficient fusion of multi-scale spectral and spatial features. The third component is a structure-aware representation layer, which combines a graph convolutional network (GCN) and a lightweight multilayer perceptron (LMLP). The GCN models global relationships in the feature map, and the LMLP further fuses multi-dimensional information, thereby improving structural modeling capability and discriminability. Finally, the model produces the final classification output through vectorization and a linear layer.

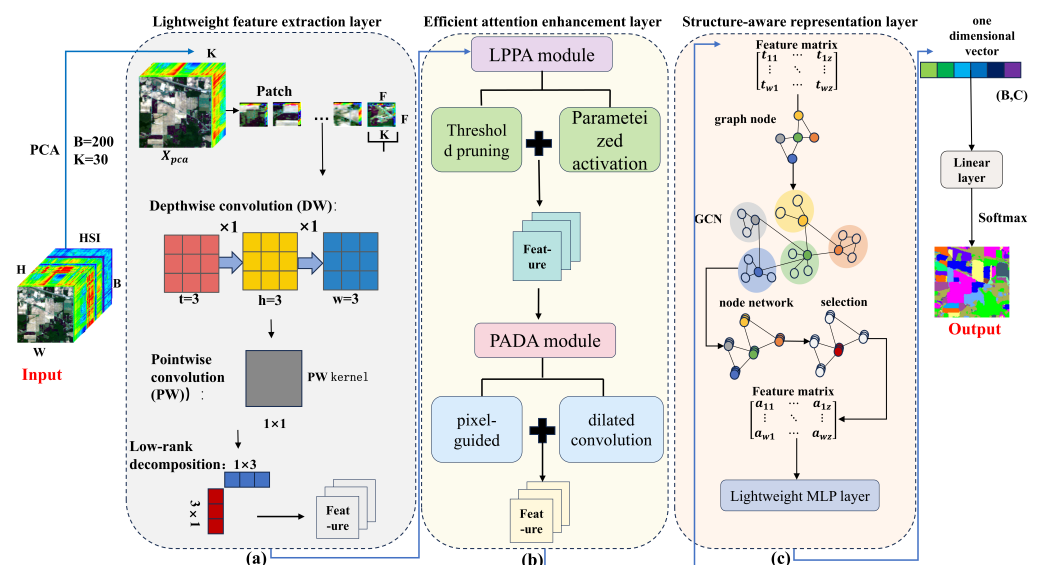


Figure 1. Overall architecture of MLSANet: (a) lightweight feature extraction layer; (b) efficient attention enhancement layer; (c) structure-aware representation layer.

3.1. Lightweight Feature Extraction Layer

In hyperspectral image classification, raw data often have very high spectral dimensionality and large spatial extent. If standard 3D convolution is directly used for feature extraction, it not only causes high computational and storage cost, but also introduces substantial redundant information. As a result, model efficiency is reduced. To reduce parameter count and computational cost while keeping feature representation capability, a lightweight feature extraction layer is proposed in this paper, as shown in Figure 2. It achieves efficient spectral–spatial feature modeling through spectral dimensionality

reduction, spatial downsampling, depthwise separable convolution, and low-rank factorization convolution.

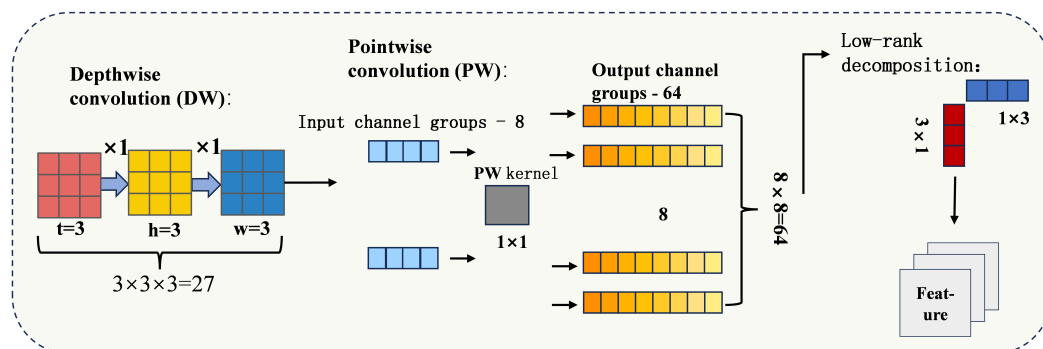


Figure 2. Lightweight feature extraction layer structure diagram.

First, PCA is applied along the spectral dimension to mitigate the strong inter-band correlation in hyperspectral data. This process effectively compresses the number of input channels while preserving the dominant spectral information, thereby reducing the computational burden of subsequent convolutional layers and, to some extent, suppressing the interference of low-variance noise components in feature extraction. Assuming an input hyperspectral image $X \in \mathbb{R}^{H \times W \times B}$, where B is the number of spectral bands, the low-dimensional representation is obtained through the PCA projection matrix $P \in \mathbb{R}^{B \times d}$, as follows:

$$X' = X \cdot P, \quad X' \in \mathbb{R}^{H \times W \times d}, \quad d \ll B \quad (1)$$

This process effectively reduces the number of input channels while preserving the dominant spectral information, thereby lowering the computational complexity and storage cost of subsequent convolution operations and helping to suppress noise interference. Next, three-dimensional average pooling (AvgPool3d) is introduced along the spatial dimension, with the kernel size set to (1, 2, 2) to perform downsampling only in the spatial domain, reducing the input from $\mathbb{R}^{H \times W}$ to $\mathbb{R}^{\frac{H}{2} \times \frac{W}{2}}$. This operation halves the spatial resolution, thereby reducing the subsequent convolution computation by approximately 75%, significantly alleviating the computational burden while retaining the main structural information.

After dimensionality reduction and downsampling, the lightweight feature extraction layer first adopts depthwise convolution (DW) to extract spatial features within each channel independently. DW convolution performs convolution operations on each channel independently, with a parameter count of $P_{DW} = C \cdot k_h \cdot k_w$ which is significantly smaller than that of standard convolution, $P_{std} = C \cdot C_{out} \cdot k_h \cdot k_w$. However, DW convolution lacks cross-channel interaction; therefore, we introduce pointwise convolution (PW) after it to achieve cross-channel feature fusion via 1×1 convolution, with a parameter count of $P_{PW} = C \cdot C_{out}$. The combination of these two operations forms depthwise separable convolution (DSC), whose total parameter count is $P_{DSC} = C \cdot k_h \cdot k_w + C \cdot C_{out}$. Compared with standard convolution, DSC reduces parameters and FLOPs by approximately 70% in the common 3×3 scenario, improving computational efficiency. Finally, to further reduce the computational redundancy in 2D convolutions, we introduce Low-rank Decomposition Convolution (LDC) on the basis of Depthwise Separable Convolution (DSC). Inspired by the concept of convolution factorization [47], this method decomposes a standard 2D convolutional kernel of size $k \times k$ into two sequential 1D kernels of sizes $k \times 1$ and $1 \times k$, thereby preserving an approximately equivalent spatial receptive field while reducing parameter counts and computational complexity. The number of parameters in a standard two-dimensional convolution can be expressed as

$$P_{2D} = C_{in} \cdot C_{out} \cdot k^2 \quad (2)$$

where C_{in} and C_{out} denote the numbers of input and output channels, respectively, and k denotes the kernel size. After applying LDC factorization, the number of parameters can be formulated as

$$P_{LDC} = C_{in} \cdot C_m \cdot k + C_m \cdot C_{out} \cdot k \quad (3)$$

where C_m represents the number of intermediate channels. As indicated by the above formulas, LDC decomposes the joint modeling process of two-dimensional spatial convolution into two consecutive one-dimensional directional convolutions, thereby reducing parameter redundancy in convolution kernels. When $k = 3$ and the intermediate channel setting is kept consistent with the input and output channels, the number of parameters can be reduced by approximately 33.3%. This factorization effectively alleviates convolutional kernel redundancy while preserving the receptive field, thus improving the lightweight property of the network.

3.2. Efficient Attention Enhancement Layer

After the lightweight feature extraction layer, the network is already capable of obtaining basic spectral and spatial features with low computational overhead. However, these features may still contain redundant information and lack sufficient discriminability. To further enhance feature representation capability, we introduce an efficient attention enhancement layer, as illustrated in Figure 3. This layer consists of two core modules: the Lightweight Pruning Parameterized Attention (LPPA) module and the Pixel-wise Adaptive Dilated Attention (PADA) module. These two modules perform attention calibration from the spectral and spatial dimensions, respectively, to enhance effective spectral responses and the representation of critical spatial regions.

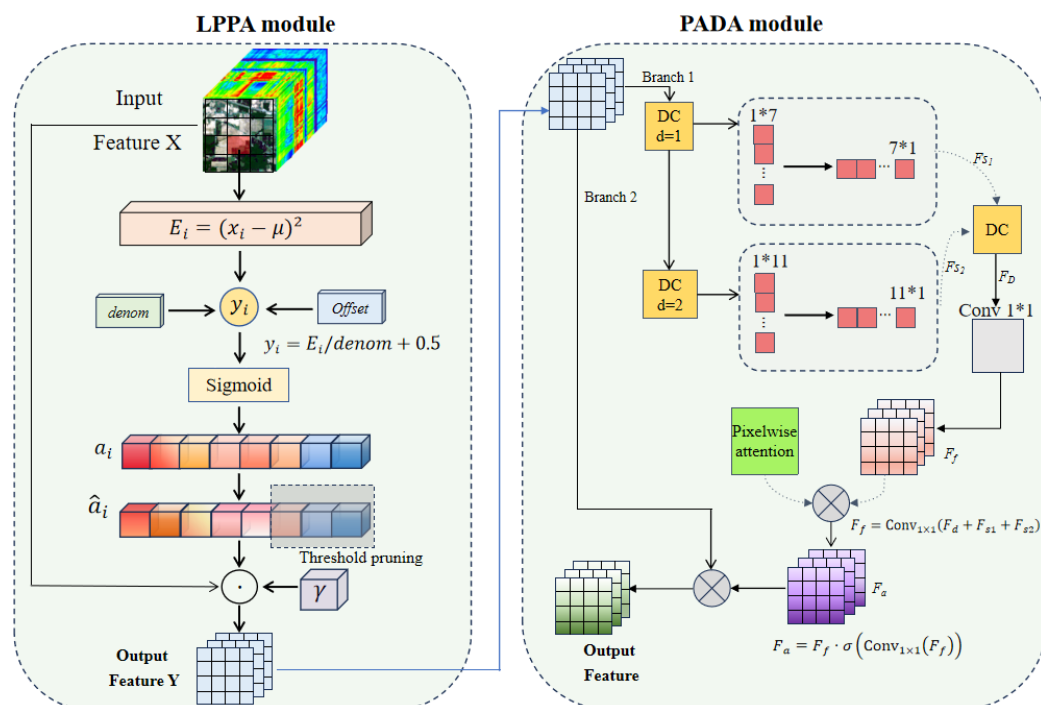


Figure 3. Structure diagram of the efficient attention enhancement layer.

3.2.1. LPPA Module

From the perspective of spectral feature modeling, hyperspectral images contain a large number of continuous and highly correlated spectral bands, among which adjacent

bands often exhibit substantial redundancy. Some low-contribution spectral responses not only fail to provide effective discriminative information but may also interfere with subsequent spatial feature modeling. To this end, LPPA is designed to screen and rescale feature responses along the spectral dimension, enabling the model to focus more on class-discriminative spectral channels while suppressing the influence of redundant spectral information. Inspired by the idea of evaluating feature importance via an energy function in SimAM [48], LPPA constructs channel-wise energy responses and incorporates a threshold-based pruning mechanism along with a learnable rescaling mechanism, thereby suppressing redundant spectral responses and enhancing effective spectral features. A comparison between SimAM and LPPA is provided in Table S3 of the Supplementary Material.

From the perspective of an energy function, the LPPA module can be viewed as a variance-based energy modeling mechanism. Specifically, the channel-wise mean of the input features is first computed, and then the squared difference between each element and its corresponding channel mean is obtained as follows:

$$E_i = (x_i - \mu)^2 \quad (4)$$

This difference can be interpreted as the “energy” at that position, reflecting the deviation of the pixel from the overall distribution. A larger energy value indicates that the position carries more significant discriminative information. To ensure scale consistency across different channels, the module adopts variance as the normalization baseline and introduces a stability term for adjustment, constructing the normalization denominator (denom). Subsequently, a stable energy score is obtained by taking the ratio of the energy to the denominator and adding a constant offset term:

$$y_i = \frac{E_i}{\text{denom}} + 0.5 \quad (5)$$

where 0.5 is an empirical offset term for the normalized energy scores, introduced to stabilize the attention responses prior to Sigmoid mapping and to prevent excessively low normalized energy values from leading to over-suppression of features. To further investigate the impact of different offset values on model performance, this paper conducts a sensitivity analysis of the offset, with the corresponding results presented in Table S5 of the Supplementary Material. The denominator term $\text{denom} = 4(\sigma^2 + \lambda)$, which denotes the energy normalization denominator, is composed of the intra-channel variance and a stability constant, and is used to perform scale normalization on the energy responses.

Then, the energy score is mapped to the (0, 1) interval using the sigmoid function to obtain the initial attention weights:

$$a_i = \sigma(y_i) \quad (6)$$

On this basis, the LPPA module further introduces energy threshold-based pruning. For pixel positions with low energy, their corresponding attention weights are directly set to zero, retaining only those salient regions with high energy:

$$\hat{a}_i = a_i \cdot \mathbf{1}_{\{a_i > \tau\}} \quad (7)$$

where τ denotes the pruning threshold, which controls the sparsity of attention weights. According to the threshold sensitivity analysis, the threshold τ is set to 0.7 in this paper.

Finally, the pruned attention weights are combined with a learnable channel scaling factor γ and multiplied element-wise with the original features to obtain the final enhanced feature representation:

$$Y = X \odot (\gamma \odot \hat{a}) \quad (8)$$

where γ denotes the learnable scaling factor, initialized to 1 and updated automatically through backpropagation during model training. It adaptively recalibrates the pruned attention weights, thereby enhancing the feature modulation capability.

3.2.2. PADA Module

After completing the lightweight attentional modeling based on the energy function, the LPPA module effectively enhances discriminative features and suppresses redundant information. However, LPPA primarily performs modeling from the perspectives of spectral response selection and channel-wise feature rescaling, and thus remains limited in capturing inter-pixel spatial relationships, local neighborhood structures, and multi-scale spatial contexts. The land cover categories in hyperspectral imagery depend not only on spectral response differences but are also closely related to spatial distribution, edge structures, and neighboring contextual information. To further strengthen the spatial context modeling capability of the model, this paper designs the Pixel-wise Adaptive Dilated Attention (PADA) module. This module expands the local receptive field through dilated convolutions, captures spatial structural information at different scales via multi-scale spatial convolutions, and adaptively modulates responses at various spatial locations using a pixel-wise attention gating mechanism, thereby enhancing the representation of key spatial regions and local contextual information.

The PADA module aims to enhance the spatial context modeling capability of features through the combination of multi-scale convolution and pixel-wise attention. This module consists of two paths: the first path directly preserves the input features as the main branch for final fusion; the second path extracts rich spatial contextual information through dilated convolution and multi-scale convolution structures. Specifically, a dilated convolution operation is first applied to the input feature x to expand the receptive field while maintaining low computational cost:

$$F_d = \text{Conv}_{3 \times 3}^{\text{dilation}=2}(x) \quad (9)$$

On this basis, two multi-scale convolution branches are designed. The first branch employs a combination of 7×1 and 1×7 convolutions, while the second branch employs a combination of 11×1 and 1×11 convolutions. These two branches are used to capture multi-scale spatial dependencies in the horizontal and vertical directions and enlarge the effective receptive field with low computational overhead. This structure enables the model to better characterize spatial distribution patterns in hyperspectral images, such as strip-like regions, block-wise areas, and continuous boundaries, while providing richer spatial contextual information for subsequent pixel-wise attention gating. Let the outputs of these two branches be F_{s1} and F_{s2} , respectively. The final fused feature is obtained by fusing them with the dilated convolution result, followed by integration via a 1×1 convolution:

$$F_f = \text{Conv}_{1 \times 1}(F_d + F_{s1} + F_{s2}) \quad (10)$$

To further enhance the response of salient regions, the PADA module introduces a pixel-wise attention branch. Specifically, a pixel-wise attention map is generated via a 1×1 convolution followed by the Sigmoid activation function, and then multiplied element-wise with the fused features to obtain the weighted feature representation:

$$F_a = F_f \cdot \sigma(\text{Conv}_{1 \times 1}(F_f)) \quad (11)$$

Finally, PADA performs pixel-level enhancement by conducting element-wise multiplication between the weighted feature F_a and the original input feature u preserved by the first path. The calculation process is formulated as

$$F_{\text{out}} = u \odot F_a \quad (12)$$

where $\odot$ denotes element-wise multiplication, and F_{out} represents the output feature of the PADA module.

The PADA module effectively integrates the global receptive field provided by dilated convolution, the directional sensitivity offered by multi-scale convolution, and the saliency modeling capability of pixel-wise attention. Compared with traditional convolution or channel attention methods, this module not only captures finer-grained spatial dependencies but also highlights regions that are more discriminative for the task, thereby improving the overall feature representation capability and model performance.

3.3. Structure-Aware Representation Layer

After the PADA module performs spatial context enhancement on the feature maps, this paper further designs a structure-aware representation layer to model the structural relationships among different spatial nodes in the form of a graph structure, while incorporating a lightweight multi-layer perceptron to achieve compact and discriminative feature representation. The core idea of this layer is to map the enhanced convolutional features into graph nodes, model the structural dependencies among nodes via a Graph Convolutional Network (GCN), and then perform feature compression and class discrimination using a Lightweight Multi-Layer Perceptron (LMLP).

Specifically, let the feature map output by PADA be denoted as $F \in \mathbb{R}^{C \times H \times W}$. First, it is transformed into a node feature matrix via reshape and flatten operations:

$$X_g \in \mathbb{R}^{N \times C} \quad (13)$$

where $N = H \times W$ denotes the number of graph nodes, with each node corresponding to a spatial location in the feature map, and C represents the feature dimension of each node. Subsequently, an adjacency matrix A is constructed based on the neighborhood relationships among nodes in the original 2D spatial domain. Specifically, if node v_i and node v_j are spatially adjacent in the 2D domain, an edge is established between them; otherwise, no connection is formed. The adjacency matrix is defined as:

$$A_{ij} = \begin{cases} 1, & v_j \in \mathcal{N}(v_i), \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

where $\mathcal{N}(v_i)$ denotes the set of spatial neighboring nodes of node v_i .

To preserve the self-information of each node and stabilize the graph convolution propagation process, we add self-connections to the adjacency matrix and applies symmetric normalization to obtain the normalized adjacency matrix as follows:

$$\hat{A} = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}} \quad (15)$$

where I denotes the identity matrix, D is the degree matrix, and its diagonal elements are defined as:

$$D_{ii} = \sum_j (A + I)_{ij} \quad (16)$$

Based on the normalized adjacency matrix, the GCN updates the node features as follows:

$$H^{(l+1)} = \sigma(\hat{A}H^{(l)}W^{(l)}) \quad (17)$$

where $H^{(l)}$ denotes the node features at the l -th layer, $W^{(l)}$ is the learnable weight matrix, and $\sigma(\cdot)$ represents the nonlinear activation function. Through graph convolutional propagation, each node not only preserves its own features but also integrates structural information from its neighboring nodes, thereby enhancing the spatial structure representation capability of the features.

A Lightweight MLP (LMLP) is further adopted to integrate and classify the structure-aware node features output by the GCN. Compared with a conventional MLP classification head, LMLP first flattens the node features generated by the GCN into a one-dimensional feature vector. Then, a low-dimensional linear projection is used to compress the flattened features into a smaller hidden feature dimension, followed by a final classification layer that maps the compact representation into the class space. Since the number of parameters in a fully connected layer is mainly determined by the product of the input and output dimensions of adjacent layers, directly classifying high-dimensional node features would introduce considerable parameter overhead. Therefore, LMLP reduces the input scale of the classification layer by first compressing the feature dimension and then performing class mapping, thereby decreasing both the parameter count and computational complexity. With this design, the structure-aware representation layer can model spatial structural relationships while generating compact and discriminative feature representations in a lightweight manner.

4. Experiments

4.1. Datasets

We use three public hyperspectral datasets to validate the proposed algorithm, namely the Indian Pines (IP) dataset, the Pavia Centre (PC) dataset, and the Salinas (SA) dataset. These datasets are described as follows.

1. The Indian Pines dataset was collected by NASA's ER-2 aircraft in 1992, covering an agricultural area in Indiana, USA. It consists of 145×145 pixels and 200 spectral bands, and is primarily used for agricultural monitoring and land cover classification research.
2. The Pavia Centre (PC) dataset was acquired by the University of Pavia, Italy, using the Reflective Optics System Imaging Spectrometer (ROSIS), covering the urban area of Pavia city center. It has a spatial resolution of 1096×715 pixels and contains 102 spectral bands, and is primarily used for urban land cover identification and classification tasks.
3. The Salinas (SA) dataset was acquired by NASA's AVIRIS sensor over the Salinas Valley in California, USA. It has a spatial resolution of 512×217 pixels and contains 204 available spectral bands, covering multiple crop types and bare soil categories. It is a commonly used benchmark dataset in agricultural monitoring and fine classification research.

Table 1 lists the detailed information of the three datasets. Tables 2–4 list the names of the land cover classes, along with the numbers of training and testing samples and the color for each class in the three datasets. Each dataset is divided into two sets: a training set and a test set. Specifically, 10%, 0.4%, and 1% of the total samples are randomly assigned to the training sets of IP, PC, and SA, respectively. The sensitivity analysis under different training proportion settings is provided in Tables S6–S8 of the Supplementary Material to validate the rationality of this partitioning strategy. This allocation ensures that all categories have sufficient samples for training, while the remaining labeled samples are used for the test set.

Table 1. Summary of the three datasets.

Dataset	Indian Pines (IP)	Pavia Centre (PC)	Salinas (SA)
Source	USDA/NASA	ROSIS	AVIRIS
Location	Southwestern Indiana, USA	Pavia, Italy	Salinas Valley, California, USA
Image size	145 × 145	1096 × 715	512 × 217
Spatial resolution	–	1.3 m	3.7 m
Number of spectral bands	224	115	224
Number of available bands	200	102	204
Wavelength range	400–2500 nm	430–860 nm	450–2450 nm
Number of land cover classes	16	9	16
Application	Agriculture, environmental monitoring	Urban remote sensing, image classification	Precision agriculture, land use classification

Table 2. Color representation, land cover classes, and numbers of training and test samples in the Indian Pines (IP) dataset.

No.	Color	Land Cover Class	Training Samples	Test Samples
01		Alfalfa	5	41
02		Corn-notill	143	1285
03		Corn-mintill	83	747
04		Corn	24	213
05		Grass-pasture	48	435
06		Grass-trees	73	657
07		Grass-pasture-mowed	3	25
08		Hay-windrowed	48	430
09		Oats	2	18
10		Soybeans-notill	97	875
11		Soybeans-mintill	245	2210
12		Soybeans-clean	59	534
13		Wheat	20	185
14		Woods	126	1139
15		Buildings-Grass-Trees-Drives	39	347
16		Stone-Steel-Towers	9	84
Total			1024	9225

Table 3. Color representation, land cover classes, and numbers of training and test samples in the Pavia Centre (PC) dataset.

No.	Color	Land Cover Class	Training Samples	Test Samples
01		Water	264	64,652
02		Trees	30	7507
03		Asphalt	12	3078
04		Brick	11	2674
05		Bitumen	26	6558
06		Tiles	37	9211
07		Shadow	29	7258
08		Meadow	171	42,655
09		Bare Soil	12	2581
Total			592	146,444

Table 4. Color representation, land cover classes, and numbers of training and test samples in the Salinas (SA) dataset.

No.	Color	Land Cover Class	Training Samples	Test Samples
01		Brocoli_green_weeds_1	20	1989
02		Brocoli_green_weeds_2	20	3706
03		Fallow	20	1956
04		Fallow_rough_plow	20	1374
05		Fallow_smooth	20	2658
06		Stubble	20	3939
07		Celery	20	3559
08		Grapes_untrained	20	11,251
09		Soil_vinyard_develop	20	6183
10		Corn_senesced_green_weeds	20	3258
11		Lettuce_romaine_4wk	20	1048
12		Lettuce_romaine_5wk	20	1907
13		Lettuce_romaine_6wk	20	896
14		Lettuce_romaine_7wk	20	1050
15		Vinyard_untrained	20	7248
16		Vinyard_vertical_trellis	20	1787
Total			320	53,809

4.2. Evaluation Metrics and Experimental Setup

4.2.1. Evaluation Metrics

Overall accuracy (OA), average accuracy (AA), and Kappa coefficient are adopted. The calculation formulas are as follows:

- (1) Overall accuracy (OA) is defined as the ratio of the count of correctly classified samples to the total sample count.

$$OA = \frac{1}{N} \sum_{i=1}^r x_{ii}$$

- (2) Average accuracy (AA) is the arithmetic mean of the classification accuracies over all classes.

$$AA = \frac{1}{C} \sum_{j=1}^C \frac{x_{jj}}{N_j}, \quad N_j = \sum_{k=1}^C x_{jk}$$

- (3) The Kappa coefficient is a measure of classification agreement that accounts for the effect of chance agreement.

$$\kappa = \frac{N \sum_{j=1}^C x_{jj} - \sum_{j=1}^C (x_{j+} x_{+j})}{N^2 - \sum_{j=1}^C (x_{j+} x_{+j})},$$

$$x_{j+} = \sum_{k=1}^C x_{jk}, \quad x_{+j} = \sum_{k=1}^C x_{kj}$$

4.2.2. Experimental Configuration

The CPU employed in the experiments is a 12th Gen Intel(R) Core(TM) i7-12700H, and the GPU is an NVIDIA GeForce RTX 3060 Laptop GPU with 13.8 GB of memory. The Adam optimizer with an adaptive learning rate is utilized to speed up convergence. The specific experimental environment is listed below: CPU: 12th Gen Intel i7-12700H; GPU: NVIDIA RTX 3060 Laptop GPU (13.8 GB); Python 3.9; PyTorch 1.12; CUDA 11.3. The optimizer is Adam with a learning rate of 0.001, the batch size is 128, the training epoch count is 100,

and the input patch size is 15×15 . The training-to-test set ratios are 10%, 0.4%, and 1% for the three datasets, respectively. The parameter settings during the experiments are detailed in Table 5.

Table 5. Experimental parameter settings.

Parameter	Setting
Learning rate	0.001
Batch size	128
Epochs	100
Patch size	15×15

4.3. Impact of Model Parameters on Classification Performance

4.3.1. Effect of Patch Size on OA

In this work, to explore the impact of various patch sizes on hyperspectral image classification performance, we analyzed the overall classification performance of the model on different datasets with patch sizes of 11×11 , 13×13 , 15×15 , 17×17 , and 19×19 , as shown in Table 6. The experimental results indicate that classification accuracy first increases and then slightly decreases as the patch size varies. This observation suggests that an appropriate patch size can achieve a favorable balance between capturing local spatial contextual information and reducing redundant information, thereby improving classification performance. In contrast, a smaller patch size (e.g., 11×11) may not fully exploit spatial information, while a larger patch size (e.g., 17×17 or 19×19), although covering more spatial context, may introduce boundary region noise and thus slightly reduce classification accuracy. Therefore, based on the experimental results across all datasets, a patch size of 15×15 is selected as the optimal choice to achieve high classification accuracy and stable feature representation.

Table 6. Effect of different patch sizes on OA.

Dataset	Patch Size				
	11×11	13×13	15×15	17×17	19×19
IP	96.10	97.35	99.00	98.39	97.68
PC	97.12	96.64	98.88	97.20	97.47
SA	96.43	97.68	98.20	97.88	98.01

4.3.2. Effect of Pruning Threshold of LPPA Module on OA

In this study, to analyze the effect of the pruning threshold in the LPPA module on classification performance, we evaluated the model's performance on different datasets with pruning thresholds set to 0.1, 0.3, 0.5, 0.7, and 0.9, as shown in Table 7. The experimental outcomes show that classification accuracy exhibits a nonlinear trend as the pruning threshold varies, with the optimal threshold differing across datasets. This observation suggests that an appropriate pruning threshold can achieve a favorable balance between preserving salient features and suppressing redundant information, thereby enhancing the discriminative capability of the model. A lower threshold (e.g., 0.1) retains more feature information but may introduce some noise, while a higher threshold (e.g., 0.9) makes the features overly sparse, discarding some useful information and thus degrading performance. Therefore, based on the experimental results across all datasets, a pruning threshold around 0.7 yields the best performance and can be adopted for the LPPA module in this study to achieve stable attention-based feature extraction.

Table 7. Effect of different pruning thresholds on OA.

Dataset	Pruning Threshold				
	0.1	0.3	0.5	0.7	0.9
IP	95.42	96.85	97.02	99.00	98.54
PC	96.38	97.14	97.99	98.88	98.75
SA	98.20	96.35	95.28	98.20	97.22

4.3.3. Impact of the Learnable Scaling Factor in the LPPA Module on OA

To further validate the effectiveness of the pruning strategy and the learnable scaling factor in the LPPA module, this study conducted comparative experiments with different combinations of pruning mechanisms, including: no pruning and no γ , pruning without γ , and pruning with γ . The OA results on each dataset are shown in Table 8. The experimental results demonstrate that the introduction of the pruning mechanism improves classification accuracy on all three datasets to varying degrees. Taking the IP dataset as an example, OA increases from 97.86% to 98.42%; for the PC dataset, from 97.53% to 98.11%; and for the SA dataset, from 97.01% to 97.63%. This indicates that appropriate attention threshold pruning can effectively suppress redundant responses and enhance salient feature expression, thereby improving the discriminative capability of the model. On this basis, further introducing the learnable scaling factor leads to additional performance gains. Specifically, the IP dataset achieves 99.00%, the PC dataset achieves 98.88%, and the SA dataset achieves 98.20%, all of which are the best results among the three mechanisms. This observation suggests that while pruning alone can improve feature sparsity, the attention response still lacks adaptive regulation capability. After introducing a lightweight learnable channel scaling factor, the model is able to adaptively regulate different feature channels, thereby enhancing important features and suppressing irrelevant information, achieving more refined attention allocation. The “pruning + γ ” mechanism achieves the best classification performance on all three datasets, validating the synergistic effectiveness of the pruning strategy and the learnable scaling factor in the LPPA module. This mechanism improves feature representation capability while maintaining low parameter overhead, demonstrating a favorable balance between lightweight design and high performance.

Table 8. Effect of different pruning mechanisms and the learnable scaling factor on OA.

Dataset	Pruning Mechanism		
	No Pruning + No γ	Pruning + No γ	Pruning + γ
IP	97.86	98.42	99.00
PC	97.53	98.11	98.88
SA	97.01	97.63	98.20

4.3.4. Effect of Dilation Rate and Kernel Size of PADA Module on OA

To investigate the effect of the dilation rate and kernel size in the PADA module on hyperspectral image classification performance, this study conducted experimental analyses on different combinations, as shown in Table 9. The experimental results indicate that classification accuracy varies significantly with different dilation rates and kernel sizes, and an appropriate combination can achieve a balance between capturing spatial contextual information and suppressing noise, thereby improving model performance. This observation suggests that a moderate dilation rate combined with a suitable kernel size can effectively expand the receptive field while preserving local spatial details, thus enhancing multi-scale feature fusion capability. In contrast, too small a dilation rate or too small a kernel size may lead to insufficient spatial context capture, whereas too large a dilation rate or too large a kernel size may introduce excessive redundant information or boundary

noise, slightly reducing classification accuracy. Therefore, based on the experimental results across all datasets, the combination of a dilation rate of 2 and a kernel size of 5×5 performs well on most datasets and can be adopted as the setting for the PADA module to achieve high accuracy and effective feature enhancement.

Table 9. Effect of different dilation rates and kernel sizes on OA.

Dataset	Dilation Rate/Kernel Size				
	1 and 3×3	2 and 3×3	2 and 5×5	3 and 5×5	2 and 7×7
IP	97.58	97.95	99.00	96.29	97.30
PC	96.23	97.82	98.88	98.74	96.24
SA	96.91	97.81	98.20	97.56	98.71

4.3.5. Effect of Pixel-Wise Attention Mechanism in PADA on OA

To validate the effectiveness of the pixel-wise attention mechanism in the PADA module, this study conducted comparative experiments with different attention types, including: no attention mechanism, the introduction of the SE channel attention mechanism, and the introduction of the pixel-wise attention mechanism. The OA results on each dataset are shown in Table 10. The experimental results show that the introduction of attention mechanisms improves the overall classification performance to varying degrees. Specifically, on the IP dataset, OA increases from 95.32% without attention to 96.38% with SE attention, and then improves to 99.00% with pixel-wise attention. On the PC dataset, OA is 96.32% without attention, slightly decreases to 95.48% with SE attention, and then increases to 98.88% with pixel-wise attention. On the SA dataset, OA is 97.62% without attention, improves to 98.00% with SE attention, and further improves to 98.20% with pixel-wise attention. These results indicate that traditional SE attention mainly performs global compression and recalibration along the channel dimension, which can enhance important spectral channels but ignores the differences between pixels in the spatial dimension. Therefore, on datasets with complex spatial structures or distinct class boundaries, its performance improvement is limited, and fine-grained spatial information may even be weakened due to global average pooling. In contrast, the pixel-wise attention mechanism independently models each spatial position, enabling the model to capture spatial structural differences more finely while preserving the integrity of channel information, thereby enhancing the discriminability between the central pixel and its neighbors. Particularly on the IP dataset, pixel-wise attention brings a significant performance gain, indicating that this mechanism has stronger adaptability in scenarios with complex class distributions and obvious spatial structural variations. In comparison, SE attention exhibits performance fluctuations on some datasets, further demonstrating that a single channel recalibration strategy cannot adequately model the spectral–spatial coupling characteristics in hyperspectral images. The pixel-wise attention mechanism achieves the best or second-best results on all three datasets, validating the effectiveness of the pixel-wise modeling strategy in the PADA module for enhancing spatial detail representation and multi-scale feature fusion, and demonstrating its stronger fine-grained modeling capability compared with traditional channel attention.

Table 10. Effect of different attention mechanisms on OA.

Dataset	Attention Mechanism		
	No Attention	SE Attention	Pixel-Wise Attention
IP	95.32	96.38	99.00
PC	96.32	95.48	98.88
SA	97.62	98.00	98.20

4.3.6. Effect of Lightweight Convolution Structure on Complexity

To validate the effectiveness of the proposed lightweight feature extraction structure in controlling model complexity, this study conducted a comparative analysis of the parameter counts of standard convolution and lightweight convolution structures, as shown in Table 11. For the 3D convolution structure, the traditional standard 3D convolution has a parameter count of 216, while the depthwise separable convolution (DW + PW) reduces the parameter count to 35, achieving a reduction of 83.8%. This result indicates that decoupling spatial convolution and channel fusion operations not only significantly reduces model complexity but also maintains the ability to effectively model spectral–spatial features. For the 2D convolution structure, the standard 3×3 convolution has a parameter count of 129,024, while the parameter count is reduced to 55,296 after adopting low-rank decomposition convolution (LDC), achieving a reduction of 57.1%. This structure decomposes a 2D convolution kernel into two 1D kernels, effectively reducing computational cost while preserving the receptive field and enhancing directional feature modeling capability. Therefore, the proposed lightweight convolution design reduces the model parameter count while preserving feature representation capability, reserving more computational budget for subsequent attention enhancement and structure-aware modeling modules, thereby achieving a favorable balance between accuracy and efficiency.

Table 11. Effect of different lightweight convolution structures on complexity.

Module Type	Convolution Type	Total Parameters	Reduction Compared with Original
3D convolution	Standard 3D convolution	216	–
	DW + PW convolution	35	83.8%
2D convolution	Standard 2D convolution	129,024	–
	LDC ($1 \times 3 + 3 \times 1$)	55,296	57.1%

4.4. Ablation Experiments

4.4.1. Ablation Analysis on Module Contributions

To verify the effectiveness of each constituent module, ablation experiments are conducted on the PC dataset, with the results presented in Table 12. The baseline denotes the fundamental architecture without the incorporation of LPPA, PADA, LMLP, GCN, or the DW + PW lightweight convolution structure. Subsequently, different modules are progressively integrated to analyze their respective impacts on classification performance.

As shown in Table 12, the individual introduction of LPPA, PADA, or LMLP improves the overall accuracy (OA) from 96.67% to over 97%, indicating that spectral response selection, spatial context enhancement, and lightweight discriminative mapping all contribute to the overall classification performance. Specifically, the use of LPPA alone leads to a slight degradation in average accuracy (AA), suggesting that spectral channel selection alone may have a certain negative impact on the average recognition performance across some categories. In contrast, both PADA and LMLP yield improvements in OA, AA, and Kappa, demonstrating that spatial context modeling and high-level feature mapping are well suited to the PC dataset.

When LPPA, PADA, and LMLP are jointly employed, the model achieves OA, AA, and Kappa of 98.45%, 95.13%, and 97.80%, respectively, representing further improvements over the single-module results and indicating a certain degree of complementarity among the spectral, spatial, and discriminative mapping modules. With the subsequent incorporation of GCN, the OA, AA, and Kappa reach 98.83%, 96.10%, and 98.35%, respectively, demonstrating that structural relationship modeling can further enhance the capability of expressing spatial correlations. Finally, the complete MLSANet, after introducing the

DW + PW lightweight convolution structure, attains the highest OA of 99.04%, along with AA and Kappa of 96.33% and 98.88%, respectively, marking the best overall performance.

Table 12. Ablation experiments on the contribution of individual modules.

Dataset	Modules					Evaluation Metrics		
	LPPA	PADA	LMLP	GCN	DW + PW	OA (%)	AA (%)	Kappa (%)
PC	×	×	×	×	×	96.67	93.55	96.57
	✓	×	×	×	×	97.17	92.87	97.14
	×	✓	×	×	×	97.06	94.44	97.10
	×	×	✓	×	×	97.15	94.16	97.12
	✓	✓	×	×	×	97.41	95.01	97.35
	✓	×	✓	×	×	98.23	94.14	97.49
	×	✓	✓	×	×	97.36	94.76	97.27
	✓	✓	✓	×	×	98.45	95.13	97.80
	✓	✓	✓	✓	×	98.83	96.10	98.35
	✓	✓	✓	×	✓	98.55	95.25	97.94
	✓	✓	✓	✓	✓	99.04	96.33	98.88

Note: ✓ indicates that the corresponding module or component is included, whereas × indicates that it is not included.

The ablation experiments demonstrate that LPPA, PADA, LMLP, GCN, and DW + PW contribute to spectral redundancy suppression, spatial context enhancement, lightweight discriminative mapping, structural relationship modeling, and low-complexity convolutional computation, respectively. The synergistic design of these modules effectively improves classification performance and enables MLSANet to achieve a favorable balance between accuracy and model complexity.

4.4.2. Ablation Study on Internal Components of LPPA

To further analyze the impact of internal components of LPPA on model performance, fine-grained ablation experiments are conducted on the PC dataset, with the results presented in Table 13. No LPPA denotes the model without any LPPA-related mechanisms, while Variance only, Pruning only, and Gamma only indicate the configurations that introduce only variance weighting, threshold-based pruning, and the learnable scaling factor γ , respectively. Full LPPA refers to the complete LPPA module that simultaneously incorporates all three mechanisms.

As shown in Table 13, compared with the No LPPA baseline, the individual introduction of any single component leads to performance improvement. Specifically, Variance only raises OA, AA, and Kappa to 98.85%, 96.33%, and 98.50%, respectively, indicating that variance-based channel response modeling contributes to enhancing discriminative spectral feature expression. Pruning only achieves the highest performance among the single-component settings, with OA, AA, and Kappa reaching 98.91%, 96.81%, and 98.60%, respectively, suggesting that threshold-based pruning can effectively suppress low-contribution spectral responses. Gamma only yields OA, AA, and Kappa of 98.74%, 96.79%, and 98.49%, respectively, demonstrating that the learnable scaling factor is capable of adaptively modulating channel responses.

When all three components are simultaneously introduced, Full LPPA achieves an OA of 99.03% and an AA of 96.82%, both higher than those of each single-component configuration, while its Kappa reaches 98.50%, remaining comparable to the single-component results. Overall, variance weighting, threshold-based pruning, and the learnable scaling factor play the roles of channel importance estimation, redundant response suppression, and response rescaling, respectively. The combination of the three mechanisms further enhances the overall classification performance.

Table 13. Ablation experiments on internal components of LPPA.

Method	Variance Weighting	Pruning	Gamma	OA (%)	AA (%)	Kappa (%)
No LPPA	×	×	×	98.59	96.07	98.29
Variance only	✓	×	×	98.85	96.33	98.50
Pruning only	×	✓	×	98.91	96.81	98.60
Gamma only	×	×	✓	98.74	96.79	98.49
Full LPPA	✓	✓	✓	99.03	96.82	98.50

Note: ✓ indicates that the corresponding module or component is included, whereas × indicates that it is not included.

4.4.3. Ablation Study on Internal Components of PADA

To further analyze the impact of internal components of PADA on model performance, fine-grained ablation experiments are conducted on the PC dataset, with the results presented in Table 14. No PADA denotes the model without any PADA-related mechanisms, while Dilated only, Multi-scale only, and Pixel attention only indicate the configurations that introduce only dilated convolution, multi-scale convolution, and pixel-wise attention gating, respectively. Full PADA refers to the complete PADA module that simultaneously incorporates all three components.

Table 14. Ablation experiments on internal components of PADA.

Method	Dilated Conv	Multi-Scale Conv	Pixel Attention	OA (%)	AA (%)	Kappa (%)
No PADA	×	×	×	98.26	96.07	98.20
Dilated only	✓	×	×	98.85	96.33	98.37
Multi-scale only	×	✓	×	98.67	95.95	98.12
Pixel attention only	×	×	✓	98.51	94.79	97.88
Full PADA	✓	✓	✓	99.02	97.05	98.61

Note: ✓ indicates that the corresponding module or component is included, whereas × indicates that it is not included.

As shown in Table 14, the individual introduction of dilated convolution raises the model's OA, AA, and Kappa to 98.85%, 96.33%, and 98.37%, respectively, all of which are improved compared with those of No PADA, indicating that expanding the spatial receptive field contributes to enhancing local context representation. In contrast, the use of multi-scale convolution or pixel-wise attention gating alone fails to consistently improve model performance. Specifically, Multi-scale only achieves an OA and Kappa of 98.67% and 98.12%, respectively, while Pixel attention only obtains OA, AA, and Kappa of 98.51%, 94.79%, and 97.88%, respectively, suggesting that spatial scale enhancement or attention modulation alone, in the absence of effective contextual modeling, may be insufficient to adequately boost classification performance.

When all three components are simultaneously introduced, Full PADA achieves an OA of 99.02%, an AA of 97.05%, and a Kappa of 98.61%, all outperforming the other configurations. This result indicates that dilated convolution, multi-scale convolution, and pixel-wise attention gating play complementary roles within PADA. Specifically, dilated convolution is employed to expand the spatial receptive field, multi-scale convolution serves to supplement local structural information at different scales, and pixel-wise attention gating adaptively modulates the fused spatial responses. The combination of the three components enables more effective enhancement of spatial context representation, thereby improving the classification performance on the PC dataset.

4.5. Comparative Experiments

To further verify the effectiveness of the proposed method, we compare MLSANet with a variety of representative hyperspectral image classification algorithms, including a conventional machine learning method (SVM), convolutional neural network-based

approaches (D3CNN [16], HybridSN [17], and SSRN [18]), hybrid-structured methods (SS-MixNet [49] and CLOLN [50]), Transformer-based methods (MorphFormer [51], CVSSN [52], GAHT [38], and SSFTT [22]), and a state space model-based method (S²Mamba [25]). Among them, HybridSN is a representative spectral–spatial hybrid network that extracts joint spectral–spatial features by combining 3D and 2D convolutions; SS-MixNet and CLOLN aim to improve classification performance from the perspectives of spectral–spatial feature interaction and lightweight feature modeling, respectively. Methods such as MorphFormer, CVSSN, GAHT, and SSFTT leverage Transformers or attention mechanisms to enhance global dependency modeling and spectral–spatial feature representation, whereas S²Mamba employs a state space model for joint spectral–spatial modeling and captures long-range dependencies in hyperspectral data through an efficient sequence modeling mechanism. Collectively, these comparison methods cover diverse technical routes, including a conventional classifier, CNNs, hybrid architectures, Transformers, and Mamba-based methods, thereby providing a comprehensive evaluation of the proposed model’s classification performance. To ensure a fair comparison, all methods are conducted under identical data split conditions, with overall accuracy (OA), average accuracy (AA), and Kappa coefficient adopted as the primary evaluation metrics. In addition, we further perform independent repeated experiments using five random seeds (0, 1, 42, 1234, and 3407) to reduce the influence of random initialization and data partitioning. The experimental results are presented in Tables 15–17. From the overall results, it can be observed that the proposed method achieves competitive classification performance on all three datasets—IP, PC, and SA—after the incorporation of the improved modules. This demonstrates that the proposed lightweight feature extraction layer and the efficient attention enhancement modules (LPPA and PADA) not only effectively enhance feature discriminability but also maintain favorable generalization capability and stability under complex scenarios.

Table 15. Classification accuracy and Kappa coefficients on the IP dataset.

No.	Traditional Method	CNN Methods			Hybrid Architectures		Transformer Methods				Mamba Method	Proposed Method
	SVM	D3CNN	HybridSN	SSRN	SS-MixNet	CLOLN	MorphFormer	CVSSN	GAHT	SSFTT	S ² Mamba	MLSANet
1	26.47	37.50	70.73	51.35	100.00	96.33	95.12	90.24	97.56	92.68	97.56	100.00 ± 0.00
2	71.81	63.92	89.50	70.08	94.79	94.71	95.80	92.04	96.97	95.57	94.86	96.61 ± 0.65
3	63.86	53.82	92.64	60.84	98.39	99.73	95.58	88.21	97.72	96.79	100.00	98.98 ± 0.88
4	43.68	30.69	90.65	44.74	98.59	98.12	97.20	92.92	94.39	98.13	100.00	100.00 ± 0.00
5	84.99	78.83	95.40	80.88	97.93	94.48	97.24	93.68	99.31	98.62	98.62	99.82 ± 0.30
6	97.19	90.66	99.70	90.58	99.54	98.93	99.70	100.00	98.48	99.85	99.24	99.97 ± 0.07
7	23.81	52.17	100.00	69.57	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00 ± 0.00
8	97.99	99.75	99.54	99.22	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00 ± 0.00
9	40.00	46.67	100.00	40.00	100.00	58.32	46.67	100.00	100.00	100.00	55.56	87.78 ± 7.24
10	73.24	64.45	89.26	70.57	99.31	98.74	96.11	94.37	92.80	96.23	98.97	98.81 ± 0.76
11	79.36	73.31	97.19	80.91	97.69	99.19	98.91	96.57	98.28	98.55	99.73	99.66 ± 0.24
12	69.52	32.08	92.32	64.84	97.19	93.26	96.07	87.59	96.44	95.88	98.31	97.57 ± 0.37
13	96.67	97.71	100.00	98.78	97.84	99.46	100.00	98.97	100.00	100.00	97.84	99.89 ± 0.24
14	92.32	93.49	98.51	93.18	100.00	99.39	99.65	98.92	99.12	99.65	100.00	100.00 ± 0.00
15	75.18	40.43	89.66	69.26	99.71	96.25	99.14	95.64	97.99	100.00	100.00	99.71 ± 0.35
16	63.24	68.75	98.81	98.67	98.81	90.48	100.00	98.86	88.10	96.43	98.81	94.29 ± 2.29
OA (%)	78.65	70.37	94.61	77.92	98.14	96.93	97.80	95.96	97.52	97.95	98.15	99.01 ± 0.16
AA (%)	68.71	64.02	93.99	73.97	98.74	78.92	94.82	95.86	97.32	98.02	96.22	98.32 ± 0.42
Kappa	75.58	65.96	93.85	74.76	97.87	96.50	97.47	95.39	97.17	97.66	97.88	98.87 ± 0.18

4.5.1. Algorithm Comparison on the IP Dataset

As shown in Table 15, MLSANet exhibits competitive classification performance compared with the other methods on this dataset. Specifically, MLSANet achieves an OA of 99.01% ± 0.16%, an AA of 98.32% ± 0.42%, and a Kappa of 98.87% ± 0.18%, yielding favorable results across the comprehensive evaluation metrics. This demonstrates that the

proposed lightweight multi-scale attention architecture can effectively enhance both the overall discriminative capability and classification stability of the model.

Table 16. Classification accuracy and Kappa coefficients on the PC dataset.

No.	Traditional Method	CNN Methods			Hybrid Architectures		Transformer Methods				Mamba Method	Proposed Method
	SVM	D3CNN	HybridSN	SSRN	SS-MixNet	CLOLN	MorphFormer	CVSSN	GAHT	SSFTT	S ² Mamba	MLSANet
1	97.91	99.94	99.19	99.86	99.93	100.00	100.00	100.00	99.97	100.00	100.00	100.00 ± 0.00
2	90.86	87.18	96.08	90.83	88.49	97.37	92.32	87.72	80.41	95.06	95.43	94.69 ± 3.31
3	64.28	97.30	92.64	92.84	96.30	87.20	92.09	92.35	93.26	93.39	91.65	92.43 ± 4.65
4	50.86	64.97	82.43	73.96	87.88	93.19	69.81	76.38	84.78	88.25	100.00	99.29 ± 0.68
5	60.61	95.09	98.43	96.52	85.24	93.92	98.86	92.64	94.98	96.43	97.45	96.16 ± 1.15
6	63.05	96.49	95.17	96.83	97.64	99.21	98.74	94.22	99.55	99.31	95.89	96.86 ± 1.18
7	36.79	87.70	87.57	87.97	96.45	97.15	90.90	84.96	81.79	84.93	97.45	97.42 ± 2.02
8	99.08	99.68	99.68	99.80	99.36	99.87	99.15	98.29	99.11	99.73	99.71	99.24 ± 0.57
9	96.92	95.96	97.59	99.41	96.63	92.98	93.00	84.88	93.14	99.20	92.46	90.52 ± 2.09
OA (%)	89.47	97.41	97.84	97.83	97.85	98.84	97.94	96.57	97.03	98.36	97.96	98.66 ± 0.25
AA (%)	73.37	91.59	94.31	93.11	94.21	95.65	92.76	90.16	91.89	95.14	96.89	96.29 ± 0.67
Kappa	96.92	96.33	96.95	96.93	96.96	98.36	97.08	95.15	95.80	97.68	97.53	98.10 ± 0.35

Table 17. Classification accuracy and Kappa coefficients on the SA dataset.

No.	Traditional Method	CNN Methods			Hybrid Architectures		Transformer Methods				Mamba Method	Proposed Method
	SVM	D3CNN	HybridSN	SSRN	SS-MixNet	CLOLN	MorphFormer	CVSSN	GAHT	SSFTT	S ² Mamba	MLSANet
1	99.15	97.99	100.00	99.70	100.00	100.00	97.80	97.10	98.05	99.55	100.00	100.00 ± 0.00
2	99.51	99.97	100.00	100.00	99.89	100.00	99.95	100.00	100.00	100.00	100.00	100.00 ± 0.00
3	89.47	91.52	96.89	99.95	98.52	99.74	88.99	99.64	98.83	98.90	100.00	100.00 ± 0.00
4	98.92	99.28	99.49	100.00	99.64	100.00	100.00	98.63	100.00	99.78	100.00	99.46 ± 0.52
5	86.60	82.99	92.11	98.68	94.70	95.07	81.42	94.07	97.19	97.69	97.63	94.27 ± 1.65
6	100.00	100.00	100.00	100.00	99.97	100.00	99.49	99.97	93.35	99.92	100.00	99.98 ± 0.01
7	97.84	98.02	99.89	100.00	99.72	99.92	99.27	99.13	99.89	99.69	97.83	99.88 ± 0.14
8	77.47	85.69	89.42	89.86	77.74	49.99	98.00	92.10	92.59	92.28	88.15	83.59 ± 3.28
9	95.38	98.21	99.92	100.00	99.82	98.14	99.98	99.87	100.00	99.98	100.00	99.83 ± 0.10
10	59.81	81.95	93.49	96.65	96.84	91.01	89.39	96.90	95.68	96.20	97.14	98.74 ± 0.50
11	66.42	75.80	87.28	86.62	100.00	91.70	88.24	99.06	98.81	88.43	98.76	99.77 ± 0.34
12	88.11	91.46	100.00	100.00	99.90	100.00	98.23	99.64	98.70	100.00	98.69	99.75 ± 0.29
13	99.45	99.23	98.90	99.01	100.00	99.89	100.00	99.67	97.15	99.56	100.00	99.62 ± 0.67
14	57.09	98.21	98.59	99.72	99.52	100.00	87.04	99.15	99.34	98.87	97.43	99.56 ± 0.62
15	51.79	46.58	84.25	96.67	96.39	98.43	25.29	91.92	90.22	94.68	97.30	98.17 ± 1.16
16	77.92	94.47	94.76	98.22	100.00	99.72	92.67	96.94	99.28	97.53	100.00	99.56 ± 0.58
OA (%)	81.91	86.16	94.28	96.82	94.28	88.14	86.62	95.40	95.00	95.24	94.90	95.88 ± 0.66
AA (%)	84.06	90.09	95.94	97.82	97.66	95.22	90.36	97.74	97.50	96.80	96.68	98.26 ± 0.33
Kappa	79.79	84.52	93.63	96.46	93.66	86.91	84.99	96.00	95.55	95.64	94.55	95.42 ± 0.73

Compared with the conventional machine learning method, MLSANet shows a clear advantage. SVM mainly relies on shallow features and is insufficient in capturing the complex spectral redundancy and spatial structural information in hyperspectral images, leading to large fluctuations in classification accuracy across multiple classes and generally low overall metrics. In contrast, MLSANet, through its lightweight feature extraction and attention enhancement mechanisms, is able to more fully exploit joint spectral–spatial features, thereby improving overall classification accuracy and consistency. Among CNN-based and hybrid-structured methods, models such as D3CNN, HybridSN, SSRN, SS-MixNet, and CLOLN already achieve favorable classification results, indicating that convolutional architectures and hybrid feature modeling approaches play a positive role in hyperspectral classification. However, these methods still exhibit certain fluctuations in classification accuracy across some categories. In contrast, MLSANet, while maintaining a low parameter count, reduces redundant computations during the initial spectral–spatial feature extraction stage through its lightweight feature extraction layer, and enhances spectral response selection and spatial context modeling via LPPA and PADA, respectively, thereby enabling the model to sustain high recognition rates across most categories. Among Transformer and attention-based methods, MorphFormer, CVSSN, GAHT, and SSFTT also achieve high accuracy, indicating that introducing global modeling mechanisms is beneficial for hyper-

spectral classification. However, such models are generally more complex in structure and incur higher computational costs. MLSANet introduces two efficient attention modules, namely LPPA and PADA, which strengthen salient feature response and spatial context modeling capability without significantly increasing computational complexity. Therefore, MLSANet still outperforms most representative methods in terms of OA, AA, and Kappa metrics, demonstrating a better balance between accuracy and efficiency. In addition, the SSM-based S^2 Mamba enhances the modeling of long-range dependencies through a joint spatial–spectral modeling mechanism, demonstrating strong feature representation capability for hyperspectral image classification. Nevertheless, such methods still rely on complex state space sequence modeling architectures, which may introduce additional computational overhead while improving global feature representation.

From the visual classification results shown in Figure 4, it can be further observed that the classification map of MLSANet is spatially smoother, exhibits higher intra-class consistency within land cover regions, has clearer class boundaries and fewer noise points, and demonstrates more stable performance especially in complex boundary areas and few-shot classes. In contrast, some compared methods suffer from obvious misclassification or speckle noise at boundaries, resulting in poor overall consistency. This indicates that the proposed method can more accurately highlight key pixel features and effectively suppress redundant information interference, thereby obtaining finer and more reliable classification results.

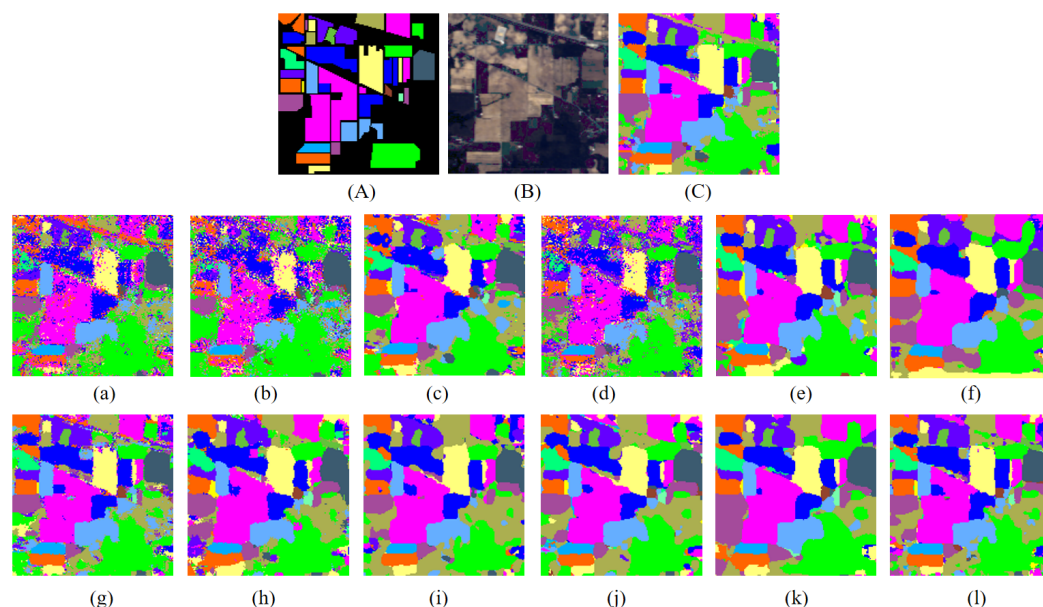


Figure 4. Classification results on the IP dataset: (A) local ground truth, (B) false color image, (C) global ground truth, (a) SVM, (b) D3CNN, (c) HybridSN, (d) SSRN, (e) SS-MixNet, (f) CLOLN, (g) MorphFormer, (h) CVSSN, (i) GAHT, (j) SSFTT, (k) S^2 Mamba, and (l) MLSANet.

4.5.2. Algorithm Comparison on the PC Dataset

From the results in Table 16, MLSANet also achieves favorable overall classification performance on this dataset, with OA, AA, and Kappa reaching $98.66\% \pm 0.25\%$, $96.29\% \pm 0.67\%$, and $98.10\% \pm 0.35\%$, respectively. These comprehensive evaluation metrics outperform those of most conventional methods, CNN-based approaches, hybrid-structured methods, and Transformer-based methods. This suggests that the proposed lightweight multi-scale attention structure maintains strong feature representation and discriminative ability while remaining compact.

Compared with the conventional machine learning method, MLSANet improves classification accuracy on all classes. SVM relies on shallow statistical features and cannot

fully capture the complex spectral redundancy and spatial structures in hyperspectral images. As a result, its accuracy on complex classes (such as classes 3, 4, and 6) is relatively low, and its overall AA is limited. In contrast, MLSANet uses its lightweight feature extraction layer and structure-aware representation mechanism to more effectively exploit joint spectral–spatial information, thereby improving the overall average accuracy and stability. Among CNN-based and hybrid-structured methods, approaches such as D3CNN, HybridSN, SSRN, SS-MixNet, and CLOLN have enhanced spectral–spatial feature representation through convolutional or hybrid feature modeling, leading to notable improvements in overall classification performance. Notably, CLOLN also achieves high OA and Kappa on this dataset, indicating that lightweight joint spectral–spatial modeling is well suited to the PC dataset. The proposed method further combines depthwise separable convolution (DW + PW), low-rank decomposition, and structural optimization. It achieves higher OA and AA than 3D-CNN and HybridSN, with a lower parameter count and computational cost, indicating a better trade-off between accuracy and efficiency. Compared with Transformer and hybrid models, methods such as MorphFormer, CVSSN, GAHT, and SSFTT achieve improved classification performance through global modeling or attention mechanisms; however, their architectures are generally more complex. In contrast, MLSANet suppresses redundant spectral responses via LPPA from the spectral dimension and enhances multi-scale contextual expression from the spatial dimension through PADA, enabling the model to maintain relatively stable recognition performance across multiple categories. Notably, on categories 4 and 7, MLSANet achieves $99.29\% \pm 0.68\%$ and $97.42\% \pm 2.02\%$, respectively, demonstrating strong class-discriminative capability. Overall, MLSANet attains a favorable trade-off between accuracy and efficiency on the PC dataset. In addition, the SSM-based S²Mamba enhances the modeling of long-range dependencies through a joint spatial–spectral modeling mechanism, achieving favorable classification performance on hyperspectral image classification tasks. However, such methods still require dedicated state space sequence modeling architectures, which may introduce additional computational overhead while improving feature representation capability.

From the qualitative visualization results shown in Figure 5, it can be further observed that the classification map of MLSANet is spatially more continuous and smoother, with fewer noise points, natural transitions in boundary regions, and reduced misclassified areas. Compared with the speckle noise and boundary blurring issues present in some CNN and Transformer methods, the proposed method can more accurately highlight key regions while maintaining intra-region consistency. This is mainly attributed to the compression of redundant information by the lightweight feature extraction layer, the enhancement of salient features by the LPPA module, and the fine-grained modeling of spatial structures by the PADA module, thereby achieving higher classification accuracy and more stable visualization performance while maintaining lightweight design.

4.5.3. Algorithm Comparison on the SA Dataset

As can be seen from Table 17, MLSANet also achieves favorable overall classification performance on the SA dataset, with OA, AA, and Kappa reaching $95.88\% \pm 0.66\%$, $98.26\% \pm 0.33\%$, and $95.42\% \pm 0.73\%$, respectively. Among all compared methods, MLSANet achieves the most competitive AA, indicating that the proposed method exhibits strong average recognition capability across different categories and maintains relatively stable classification performance even under imbalanced class distributions.

Compared with the conventional machine learning method, the performance gap is evident. SVM has large accuracy fluctuations on complex classes such as classes 5, 8, 11, 15, and 16, suggesting that it cannot effectively model the complex spectral redundancy and spatial structure in hyperspectral data. In contrast, MLSANet is capable of more thor-

oughly exploiting spectral–spatial joint features through its lightweight feature extraction layer, spectral redundancy suppression via LPPA, and spatial context enhancement via PADA, thereby improving the average classification performance across different categories. Among CNN-based and hybrid-structured methods, approaches such as D3CNN, HybridSN, SSRN, SS-MixNet, and CLOLN enhance spectral–spatial feature representation through convolutional modeling or hybrid feature extraction, achieving substantially improved overall classification performance over conventional methods. Notably, SSRN achieves high OA and Kappa on the SA dataset, indicating that deep residual spectral–spatial modeling is well adapted to this dataset. Although MLSANet does not yield the highest OA and Kappa, it attains an AA of $98.26\% \pm 0.33\%$, which is higher than those of most compared methods, suggesting that the proposed method exhibits favorable stability in terms of average per-class recognition capability. In comparison with Transformer-based and attention-based methods, MorphFormer, CVSSN, GAHT, and SSFTT improve classification performance through global modeling or attention mechanisms, yet still exhibit certain fluctuations across different categories. MLSANet, by suppressing redundant spectral responses via LPPA from the spectral dimension and enhancing multi-scale contextual expression from the spatial dimension through PADA, maintains high recognition rates on relatively challenging categories such as classes 3, 10, 11, 15, and 16. In particular, on classes 11 and 15, MLSANet achieves $99.77\% \pm 0.34\%$ and $98.17\% \pm 1.16\%$, respectively, demonstrating that the proposed modules possess strong discriminative capability for easily confusable categories. In addition, the SSM-based S^2 Mamba enhances the modeling of long-range dependencies through a joint spatial–spectral modeling mechanism. However, its state space architecture may still introduce additional computational overhead.

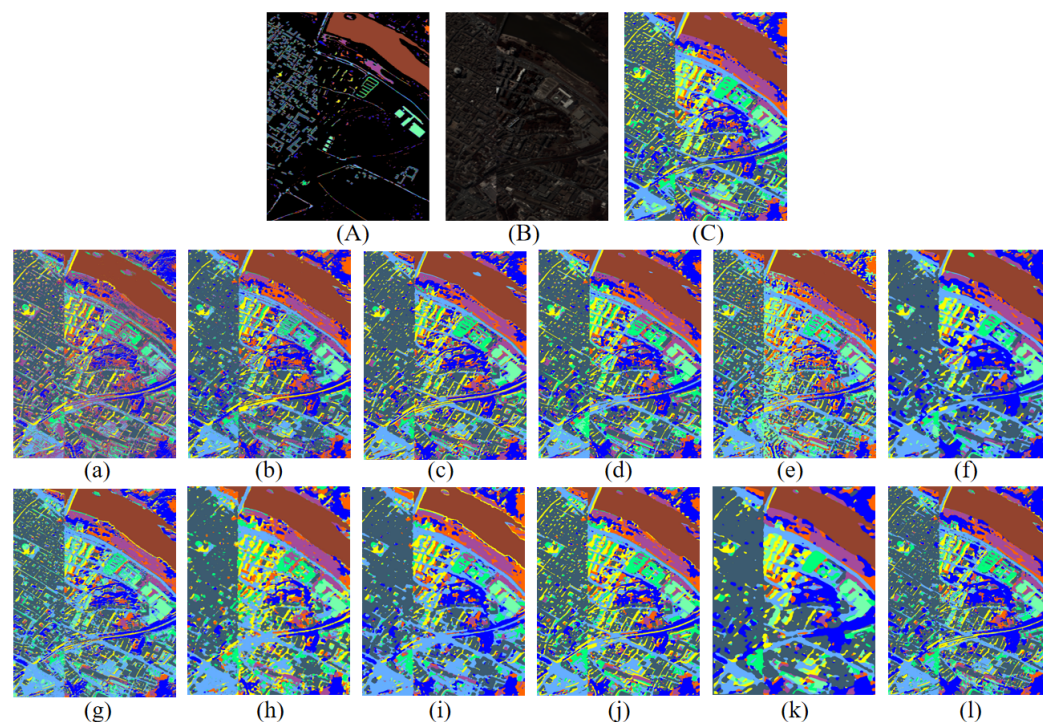


Figure 5. Classification results on the PC dataset: (A) local ground truth, (B) false color image, (C) global ground truth, (a) SVM, (b) D3CNN, (c) HybridSN, (d) SSRN, (e) SS-MixNet, (f) CLOLN, (g) MorphFormer, (h) CVSSN, (i) GAHT, (j) SSFTT, (k) S^2 Mamba, and (l) MLSANet.

From the visualization results shown in Figure 6, it can be seen that MLSANet maintains high recognition integrity even in small-sample regions and on fine land covers, without evident over-smoothing or structural loss. Compared with the “blocky” appear-

ance or detail blurring that some methods tend to produce in complex texture regions, the proposed model better preserves the spatial structure of the original scene, indicating improved fine-grained feature representation capability.

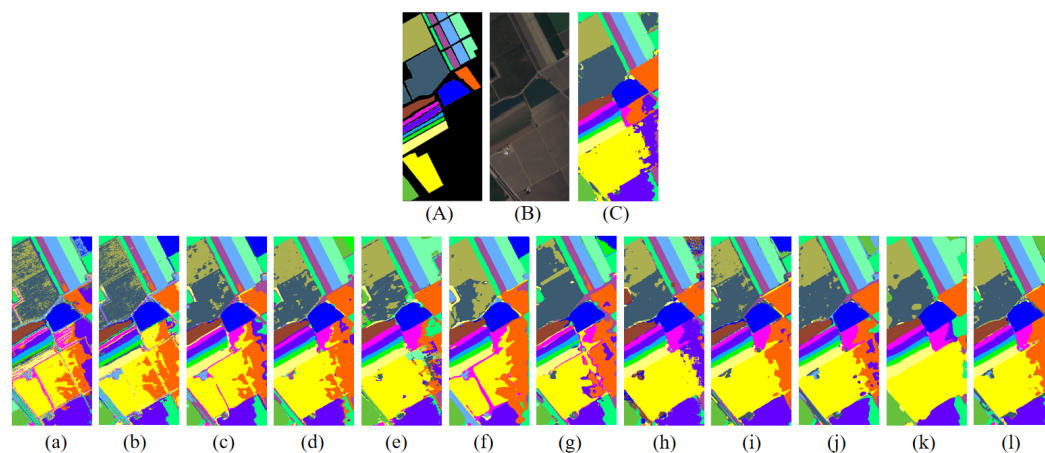


Figure 6. Classification results on the SA dataset: (A) local ground truth, (B) false color image, (C) global ground truth, (a) SVM, (b) D3CNN, (c) HybridSN, (d) SSRN, (e) SS-MixNet, (f) CLOLN, (g) MorphFormer, (h) CVSSN, (i) GAHT, (j) SSFTT, (k) S²Mamba, and (l) MLSANet.

4.6. Computational Complexity Analysis

To further verify the lightweight characteristics of MLSANet, we evaluate the model complexity at the full-network level and compare it with several representative hyperspectral image classification methods. The results are presented in Table 18. The evaluation metrics include parameter count, FLOPs, inference time, and GPU memory consumption. In contrast to studies that compare only the complexity of individual convolutional operators, Table 18 reports the overall computational cost of a complete forward pass of the entire model, covering all major modules, including the lightweight feature extraction layer, LPPA, PADA, the GCN-based structure-aware layer, and LMLP. Therefore, these statistics provide a more comprehensive reflection of the practical computational complexity of the model.

As shown in Table 18, the proposed method exhibits clear lightweight characteristics and high computational efficiency across all three datasets. In terms of parameter count, MLSANet achieves only 105.78 K, 30.25 K, and 110.52 K on the IP, PC, and SA datasets, respectively, all of which are lower than those of most CNN-based, Transformer-based, hybrid-structured, and Mamba-based hyperspectral classification models. In comparison with methods such as HybridSN, GAHT, SS-MixNet, S²Mamba, and SSFTT, MLSANet demonstrates a considerably reduced parameter scale, reflecting a more compact architectural design.

In terms of FLOPs, MLSANet achieves 8.90 M, 7.32 M, and 6.58 M on the three datasets, respectively, which are also lower than those of HybridSN, CVSSN, GAHT, SS-MixNet, S²Mamba, and SSFTT, indicating that modules such as PADA and GCN do not introduce excessive overall computational burden. It should be noted that the FLOPs reported in this paper correspond to the total computational cost of a complete forward pass of the entire MLSANet model. Therefore, these results provide a more complete reflection of the actual computational overhead of the model.

In terms of inference time and GPU memory consumption, MLSANet also demonstrates favorable resource utilization efficiency while maintaining low parameter count and computational complexity. Overall, by leveraging lightweight feature extraction, parameter optimization, and an efficient attention mechanism, MLSANet effectively re-

duces model complexity while preserving superior classification performance. Compared with conventional CNN-based, Transformer-based, and advanced Mamba-based methods such as S²Mamba, MLSANet achieves a favorable trade-off between accuracy and efficiency, making it more suitable for hyperspectral image classification under resource-constrained scenarios.

Table 18. Performance comparison of representative models on three hyperspectral image datasets.

Dataset	Metric	D3CNN	HybridSN	MorphFormer	CVSSN	GAHT	SpectralFormer	SS-MixNet	S ² Mamba	SSFTT	MLSANet
IP	Parameters/K	163.28	540.23	200.69	348.96	846.79	141.34	327.4	128.52	672.96	105.78
	FLOPs/M	279.82	254.32	18.45	110.55	160.32	4.44	542.51	16.23	280.26	8.90
	Time/s	8	4	4	5	2	2.65	23.91	0.33	10	2
	GPU Memory/MB	128.27	104.72	463.67	204.3	554.93	48.24	856.42	365.69	64.71	176.78
PC	Parameters/K	95.42	324.58	150.42	301.24	900.06	140.88	379.75	135.62	200.62	30.25
	FLOPs/M	70.26	168.32	28.33	95.33	46.32	4.44	696.81	20.85	28.54	7.32
	Time/s	23	20	40	18	18	2.68	30.25	5.45	12	20
	GPU Memory/MB	124.67	101.37	464.44	204.89	556.06	48.78	1107.56	467.05	65.05	183.99
SA	Parameters/K	280.42	523.89	198.8	360.92	983.04	141.34	327.4	128.52	1298.72	110.52
	FLOPs/M	134.52	250.50	20.21	105.70	63.88	4.44	542.51	16.23	76.74	6.58
	Time/s	10	20	25	24	12	2.8	24.03	1.54	20	11
	GPU Memory/MB	133	110.44	460.01	200.11	551.27	48.24	866.42	365.69	61.06	180.01

5. Discussion

The experimental results demonstrate that MLSANet achieves a favorable balance between classification accuracy and computational efficiency on three public hyperspectral image datasets. Compared with conventional machine learning methods and representative deep learning models, the proposed method maintains competitive classification performance while using a relatively small parameter scale and low FLOPs. This performance is mainly attributed to the coordinated design of the lightweight feature extraction layer, LPPA, PADA, and the structure-aware representation layer. Specifically, the lightweight feature extraction layer reduces redundant computation in the early stage, LPPA suppresses low-contribution spectral responses through pruning and learnable scaling, PADA strengthens multi-scale spatial contextual representation, and the GCN-based structure-aware layer further models relationships among enhanced features. These components jointly improve the discriminability of spectral–spatial features without introducing excessive computational overhead.

Despite the promising results achieved by the proposed method, there remains room for further improvement. First, this work primarily focuses on lightweight spectral–spatial feature enhancement under a fixed data partition setting, and the generalization capability under cross-dataset, cross-sensor, and cross-scene conditions has yet to be thoroughly investigated. Future work may incorporate domain adaptation, transfer learning, or cross-domain feature alignment strategies to further enhance the transferability of the model across different hyperspectral scenarios. Second, the pruning threshold, attention weights, and multi-scale spatial modeling schemes in LPPA and PADA currently rely on pre-defined structures. Further research could explore adaptive threshold selection, dynamic convolution, or variable-scale attention mechanisms to enable the model to automatically adjust its feature enhancement behavior according to different land-cover categories and spatial structures. Finally, although MLSANet maintains relatively low complexity in terms of parameter count and FLOPs, further consideration of model compression, quantization, knowledge distillation, and hardware-friendly deployment strategies is still required for edge-device or real-time remote sensing applications so as to improve the practical inference efficiency.

6. Conclusions

In this paper, a Multi-Scale Lightweight Spectral Attention Network (MLSANet) is proposed for hyperspectral image classification. The proposed method aims to reduce model complexity while improving spectral–spatial feature representation. By combining lightweight feature extraction, spectral response pruning and recalibration, pixel-wise adaptive dilated attention, and structure-aware representation, MLSANet effectively suppresses redundant information and enhances discriminative feature learning. Experiments on the Indian Pines, Pavia Centre, and Salinas datasets show that MLSANet achieves competitive classification accuracy with relatively low computational complexity, demonstrating its potential for efficient hyperspectral image classification in resource-constrained scenarios.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/rs18152451/s1>, Figure S1. Attention heatmap of the LPPA spectral response pruning module; Figure S2. Pixel-wise attention heatmap of the PADA module; Figure S3. Confusion matrix of the baseline method on the Indian Pines dataset (IP); Figure S4. Confusion matrix of the full model on the Indian Pines dataset (IP); Figure S5. Confusion matrix of the model without the LMLP module on the Indian Pines dataset (IP); Figure S6. Confusion matrix of the model without the LPPA module on the Indian Pines dataset (IP); Figure S7. Confusion matrix of the model without the PADA module on the Indian Pines dataset (IP); Table S1. Comparative Analysis of Inference Efficiency; Table S2. Impact of Lightweight Architecture on Classification Performance; Table S3. Effectiveness evaluation of LPPA components compared with SimAM (IP); Table S4. Channel masking analysis between LPPA scores and classification performance; Table S5. Sensitivity analysis of the offset term in LPPA; Table S6. Sensitivity analysis of training sample proportions on the IP dataset; Table S7. Sensitivity analysis of training sample proportions on the PC dataset; Table S8. Sensitivity analysis of training sample proportions on the SA dataset.

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

Funding: This research was funded by the Natural Science Foundation of Liaoning Province, grant number 2024-MS-168; the Fundamental Research Funds for the Provincial Universities of Liaoning, grant number LJ212410150030; and the Open Fund of the National Key Laboratory of Maritime Technology and Safety, grant number SKLMTS-DMU-2025-04. The APC was funded by Yanjuan Wang.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments.

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

References

- Goetz, A.F.; Vane, G.; Solomon, J.E.; Rock, B.N. Imaging spectrometry for earth remote sensing. *Science* **1985**, *228*, 1147–1153. [[CrossRef](#)] [[PubMed](#)]
- Chang, C.I. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*; Springer Science & Business Media: New York, NY, USA, 2003; Volume 1.
- Bioucas-Dias, J.M.; Plaza, A.; Camps-Valls, G.; Scheunders, P.; Nasrabadi, N.; Chanussot, J. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–36. [[CrossRef](#)]
- Singh, R.; Kumar, V. Evaluating automated endmember extraction for classifying hyperspectral data and deriving spectral parameters for monitoring forest vegetation health. *Environ. Monit. Assess.* **2023**, *195*, 72.

5. Shimoni, M.; Haelterman, R.; Perneel, C. Hyperspectral imaging for military and security applications: Combining myriad processing and sensing techniques. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 101–117. [\[CrossRef\]](#)
6. de Cássia Mariotti, K.; Ortiz, R.S.; Ferrão, M.F. Hyperspectral imaging in forensic science: An overview of major application areas. *Sci. Justice* **2023**, *63*, 387–395. [\[CrossRef\]](#)
7. Song, Y.; Zhang, J.; Liu, Z.; Xu, Y.; Quan, S.; Sun, L.; Bi, J.; Wang, X. Deep learning for hyperspectral image classification: A comprehensive review and future predictions. *Inf. Fusion* **2025**, *123*, 103285. [\[CrossRef\]](#)
8. Zhao, G.; Zheng, X.; Zhu, L.; Chen, G.; Guo, W.; Chen, Z.; Lu, Y. Detection of defects in Micro-LED arrays by AOTF hyperspectral imaging and machine learning. *Opt. Laser Technol.* **2025**, *184*, 112555. [\[CrossRef\]](#)
9. Felegari, S.; Sharifi, A.; Khosravi, M.; Sabanov, S. Using experimental models and multitemporal Landsat-9 images for cadmium concentration mapping. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 2503204. [\[CrossRef\]](#)
10. Yuen, P.W.; Richardson, M. An introduction to hyperspectral imaging and its application for security, surveillance and target acquisition. *Imaging Sci. J.* **2010**, *58*, 241–253. [\[CrossRef\]](#)
11. Ahmad, M.; Distefano, S.; Khan, A.M.; Mazzara, M.; Li, C.; Li, H.; Aryal, J.; Ding, Y.; Vivone, G.; Hong, D. A comprehensive survey for hyperspectral image classification: The evolution from conventional to transformers and mamba models. *Neurocomputing* **2025**, *644*, 130428. [\[CrossRef\]](#)
12. Paoletti, M.E.; Haut, J.M.; Plaza, J.; Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 279–317. [\[CrossRef\]](#)
13. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep learning-based classification of hyperspectral data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2094–2107. [\[CrossRef\]](#)
14. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep convolutional neural networks for hyperspectral image classification. *J. Sens.* **2015**, *2015*, 258619. [\[CrossRef\]](#)
15. Zhao, W.; Du, S. Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4544–4554. [\[CrossRef\]](#)
16. 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. [\[CrossRef\]](#)
17. Roy, S.K.; Krishna, G.; Dubey, S.R.; Chaudhuri, B.B. HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 277–281.
18. Zhong, Z.; Li, J.; Luo, Z.; Chapman, M. Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 847–858. [\[CrossRef\]](#)
19. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [\[CrossRef\]](#)
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
21. Zhang, G.; Abdulla, W. Transformers Meet Hyperspectral Imaging: A Comprehensive Study of Models, Challenges and Open Problems. *Eng. Appl. Artif. Intell.* **2026**, *169*, 113947. [\[CrossRef\]](#)
22. 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. [\[CrossRef\]](#)
23. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752.
24. Huang, L.; Chen, Y.; He, X. Spectral–Spatial Mamba for Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 2449. [\[CrossRef\]](#)
25. Wang, G.; Zhang, X.; Peng, Z.; Zhang, T.; Jiao, L. S²Mamba: A Spatial–Spectral State Space Model for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5511413. [\[CrossRef\]](#)
26. Xu, F.; Mei, S.; Zhang, Y.; Zhan, D.; Zeng, L. Hierarchical Spatial Spectral Mamba for Hyperspectral Image Classification. In *Proceedings of the IGARSS 2025–2025 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: New York, NY, USA, 2025; pp. 8431–8434.
27. Zhang, H.; Xu, X.; Li, S.; Plaza, A. Wavelet Decomposition-Based Spectral–Spatial Mamba Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5518817. [\[CrossRef\]](#)
28. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 3–19.
30. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 11534–11542.
31. Miao, T.; Zeng, H.; Yang, W.; Chu, B.; Zou, F.; Ren, W.; Chen, J. An improved lightweight RetinaNet for ship detection in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4667–4679. [\[CrossRef\]](#)

32. Yang, S.; An, W.; Li, S.; Wei, G.; Zou, B. An improved FCOS method for ship detection in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8910–8927. [\[CrossRef\]](#)
33. Liu, S.; Li, D.; Jiang, R.; Liu, Q.; Wan, J.; Yang, X.; Liu, H. A mixed-scale self-distillation network for accurate ship detection in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 9843–9857. [\[CrossRef\]](#)
34. Bai, L.; Yao, C.; Ye, Z.; Xue, D.; Lin, X.; Hui, M. Feature enhancement pyramid and shallow feature reconstruction network for SAR ship detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1042–1056. [\[CrossRef\]](#)
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
36. He, J.; Zhao, L.; Yang, H.; Zhang, M.; Li, W. HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 165–178. [\[CrossRef\]](#)
37. 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.* **2021**, *60*, 5518615. [\[CrossRef\]](#)
38. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5539014. [\[CrossRef\]](#)
39. Yang, X.; Cao, W.; Lu, Y.; Zhou, Y. Hyperspectral image transformer classification networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5528715. [\[CrossRef\]](#)
40. Zhong, Z.; Li, Y.; Ma, L.; Li, J.; Zheng, W.S. Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5514715. [\[CrossRef\]](#)
41. Hang, R.; Li, Z.; Liu, Q.; Ghamisi, P.; Bhattacharyya, S.S. Hyperspectral Image Classification with Attention-Aided CNNs. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 2281–2293. [\[CrossRef\]](#)
42. Huang, W.; Zhao, Z.; Sun, L.; Ju, M. Dual-Branch Attention-Assisted CNN for Hyperspectral Image Classification. *Remote Sens.* **2022**, *14*, 6158. [\[CrossRef\]](#)
43. Liang, H.; Bao, W.; Shen, X.; Zhang, X. HSI-Mixer: Hyperspectral Image Classification Using the Spectral-Spatial Mixer Representation From Convolutions. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6013005. [\[CrossRef\]](#)
44. Li, G.; Ye, M. Hyperspectral Image Classification via Transformer-Based Spectral-Spatial Attention Decoupling and Adaptive Gating. *Int. J. Image Data Fusion* **2026**, *17*, 2618653. [\[CrossRef\]](#)
45. Ahmad, I.; Farooque, G.; Hadi, F.; Khader, A.; Ghorashi, S.A.; Ahmed, A.; Alabdulkreem, E. A Multiscale Transformer with Spatial Attention for Hyperspectral Image Classification. *Sci. Rep.* **2026**, *16*, 4690. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Mahdavi-pour, Z.; Alromema, N.; Khader, A.; Farooque, G.; Ahmed, A.; Damos, M.A. TGDHTL: Hyperspectral Image Classification via Transformer-Graph Convolutional Network-Diffusion with Hybrid Domain Adaptation. *Remote Sens.* **2026**, *18*, 189. [\[CrossRef\]](#)
47. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 2818–2826.
48. Yang, L.; Zhang, R.; Li, L.; Xie, X. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, Online, 18–24 July 2021.
49. Alkhatib, M.Q. Hyperspectral Image Classification Using Spectral-Spatial Mixer Network. *arXiv* **2025**, arXiv:2511.15692.
50. Li, C.; Rasti, B.; Tang, X.; Duan, P.; Li, J.; Peng, Y. Channel-Layer-Oriented Lightweight Spectral-Spatial Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5504214. [\[CrossRef\]](#)
51. 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. [\[CrossRef\]](#)
52. Li, M.; Liu, Y.; Xue, G.; Huang, Y.; Yang, G. Exploring the Relationship Between Center and Neighborhoods: Central Vector Oriented Self-Similarity Network for Hyperspectral Image Classification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 1979–1993. [\[CrossRef\]](#)

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.