Integrating Multiscale Spatial–Spectral Shuffling Convolution With 3-D Lightweight Transformer for Hyperspectral Image Classification

Qinggang Wu, Mengkun He, Qiqiang Chen, Le Sun, *Member, IEEE*, and Chao Ma

***Abstract*—The combination of convolutional neural networks and vision transformers has garnered considerable attention in hyperspectral image (HSI) classification due to their abilities to enhance the classification accuracy by concurrently extracting local and global features. However, these accuracy improvements come at the cost of significant demands on storage resources, computational overhead, and extensive training samples. To address these challenges, this article proposes a multiscale spatial–spectral shuffling convolution integrated with a 3-D lightweight transformer (MSC-3DLT) for HSI classification. This network directly captures 3-D structural features throughout the entire feature extraction process, thereby enhancing HSI classification performance even at small sampling rates within a lightweight framework. Specifically, we first design a multiscale spatial–spectral shuffling convolution to comprehensively refine spatial–spectral feature granularities and enhance feature interactions by shuffling multiscale features across different groups. Second, to maximize the exploitation of limited training samples, we rethink transformers from the 3-D structural perspective of HSI data and propose a novel 3-D lightweight transformer (3DLT). Different from the slicing operation employed in classical transformers, the 3DLT directly extracts the inherent 3-D structural features from the HSI and mitigates quadratic complexity through a lightweight spatial–spectral pooling cross-attention mechanism. Finally, a novel training strategy is designed to adaptively adjust the learning rate based on multimetric feedback during the model training process, significantly accelerating the model fitting speed. Extensive experiments demonstrate that the proposed MSC-3DLT method remains highly competitive compared with state-of-the-art methods in terms of classification accuracy, model parameters, and floating point and operations under small sampling rates.**

***Index Terms*—Convolutional neural networks (CNNs), hyperspectral image (HSI) classification, multimetric adaptive learning rate (MALR), 3-D lightweight transformer (3DLT).**

I. Introduction

A HYPERSPECTRAL image (HSI) captures abundant spectral information of ground materials across multiple narrowbands [1]. It effectively recognizes internal object features to classify HSIs in various fields, such as land cover identification [2], [3], precision agriculture [4], [5], environmental monitoring [6], [7], and military reconnaissance [8], [9].

In the past, traditional HSI classification methods, such as support vector machine [10], [11], *K*-nearest neighbors [12], [13], principal component analysis (PCA) [14], [15], linear discriminant analysis [16], [17], etc., depended on manually designed feature extraction and classification methods. These methods are not robust to complex environments and data distributions for the requirement of tuning hyperparameters, which hinders the improvement of HSI classification performance.

As the rapid development of the deep learning technique, the convolutional neural network (CNN) becomes prevalent in HSI classification by utilizing abundant spatial and spectral features and achieves superior performance [18]. 1-D CNN methods extract spectral features from multiple regions in parallel [19], [20], [21]. 2-D CNN methods extract intuitive spatial features of edges and textures, which provides comprehensive feature recognition ability [22], [23], [24]. To address the problem that the same materials exhibit different spectral signatures and vice versa, 3-D CNN methods maximize the ability to extract 3-D structure features by directly processing the cubic HSI data [25], [26], [27]. In addition, multiscale feature extraction captures diverse information from various receptive fields, which includes small-scale detailed texture features as well as large-scale semantic object contour features. Moreover, it reduces noise impact on classification results by introducing spatial contextual information from different scales, which increases model generalization ability [28], [29], [30]. However, the huge feature dimension may increase the complexity of feature fusion.

The attention mechanism (AM) has been widely applied to concentrate on important HSI features, which substantially improve classification performance. Traditional AMs increase the utilization of HSI data by enhancing the spatial and spectral features [31], [32], [33]. Recently, the transformer has attracted

Received 30 October 2024; revised 31 December 2024; accepted 8 January 2025. Date of publication 23 January 2025; date of current version 17 February 2025. This work was supported in part by the Young Backbone Teacher Training Program of Henan Province under Grant 2023GGJS090, in part by the Scientific and Technological Research Project of Henan Provincial Department of Science and Technology under Grant 242102210013, Grant 242102210017, and Grant 232102211048, in part by the National Natural Science Foundation of China under Grant 61502435 and Grant 62476255, and in part by the Key Scientific Research Projects of Colleges in Henan Province under Grant 23A520001. *(Corresponding author: Qinggang Wu.)*

Qinggang Wu, Mengkun He, Qiqiang Chen, and Chao Ma are with the College of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450000, China (e-mail: wuqg@zzuli.edu.cn; 332207050644@email.zzuli.edu.cn; 402895213@qq.com; 208260257@qq.com).

Le Sun is with the School of Computer Science and the Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: sunlecncom@163.com).

The code of this work is available online at https://github.com/H892328874/MSC-3DLT.

Digital Object Identifier 10.1109/JSTARS.2025.3533211

widespread attention from researchers in image processing field due to its ability to extract long-range dependencies through the self-attention mechanism. Hong et al. [34] proposed a SpectralFormer that first applied the vision transformer (ViT) in HSI classification by grouping and stacking spectral tokens to obtain global spectral information with self-attention. Although the ViT excels at extracting long-range spectral sequential features, it neglects the local spatial contextual features.

It has been widely acknowledged that the combination of CNN and ViT substantially improves HSI classification performance [35]. Zhang et al. [36] proposed a convolution transformer mixer with a dual-branch parallel network architecture to extract local and global features through the CNN and the Transformer, respectively. Mei et al. [37] introduced a group-aware hierarchical transformer (GAHT) to improve classification accuracy by enforcing local–global spectral relationships through group convolution and transformer encoder blocks. Yu et al. [38] developed a HyperSINet by fusing Conv2Trans and Trans2Conv in a soft-selective weighted manner to facilitate mutual complementation between local and nonlocal HSI features. Sun et al. [39] presented a spectral–spatial feature tokenization transformer (SSFTT) by incorporating the CNN to extract low-level spatial–spectral information, which is weighted by a Gaussian function and effectively represented by a transformer encoder. Roy et al. [40] proposed a morphological transformer (morphFormer), which improves the AMs of the ViT by employing spectral–spatial morphological convolution to reduce the number of parameters. Although the combination of CNN and Transformer achieves promising HSI classification results, there are still several challenges. First, the quadratic complexity of Transformer increases computational overhead and storage resource requirements [41]. Second, it is essential to improve the utilization of HSI features at smaller sampling rates for limited HSI samples [42]. Third, it is crucial to minimize time consumption and resource requirements during the model training process.

To address these problems, we propose a multiscale spatial–spectral shuffling convolution integrated with a 3-D lightweight transformer (MSC-3DLT) for HSI classification. To improve the utilization of HSI features with limited samples, the multiscale spatial–spectral feature shuffling convolution (MSC) module is designed to extract spatial–spectral features at different granularities, shuffle the multiscale features across different groups, and enhance spectral features. In addition, the 3-D lightweight transformer (3DLT) module, based on the lightweight spatial–spectral pooling cross-attention (LSPCA) mechanism, is proposed to extract and aggregate both smooth and salient features while reducing computational overhead. Finally, to accelerate the model training speed, a multimetric adaptive learning rate (MALR) scheduler is presented to dynamically adjust the learning rate according to the model state during the training process. Overall, the proposed MSC-3DLT method effectively balances model performance, storage resources, and computation overhead in HSI classification. The contributions of this article are summarized as follows.

1) We propose a 3DLT module to preserve the spatial structure in the HSI by designing an LSPCA mechanism to replace the self-attention mechanism, thereby reducing quadratic complexity.
2) We design a novel MSC module to enhance feature utilization by shuffling the multigranularity features, which improves information flow and interaction.
3) A new training strategy is proposed based on the MALR scheduler, which adaptively adjusts the learning rate by continuously monitoring multiple feedback during the training process.
4) Extensive experiments validate the effectiveness of the proposed method, demonstrating state-of-the-art (SOTA) classification performance on the HSI under small sampling rates, with fewer parameters and floating point and operations (FLOPs).

The rest of this article is organized as follows. Section II introduces related techniques and works. Section III presents the details of the proposed MSC-3DLT method. Extensive experiments and comparisons are conducted in Section IV. Section V discusses the limitations and future research directions. Finally, Section VI concludes this article.

II. Related Work

A. Lightweight ViT for HSI Classification

Transformer has achieved a great success in natural language processing (NLP) tasks [43] due to its ability of capturing long-range dependencies via self-attention mechanisms. Dosovitskiy et al. [44] adapted it to the ViT for the first time to classify natural images by splitting them into a sequence of tokens and utilize self-attention to capture global spatial dependencies. Afterward, the ViT attracts much attention from researchers in image processing field.

Generally, Transformer includes three steps: patch embedding, self-attention, and multilayer perceptron (MLP). For HSI classification, it is first split into different tokens of $[T^1, T^2, \ldots, T^i]$ and concatenated with the ClassToken T_{cls} for final classification. The position information of each token is denoted by position encoding PE, and represented as follows:

$$T_{\text{in}} = \left[T_{\text{cls}}, T^1, T^2, \ldots, T^i\right] + \text{PE}. \tag{1}$$

In the second step, the input T_{in} is subsequently fed to Transformer to extract global features. As the most important component, the self-attention mechanism effectively captures the dependencies between spectral sequences. Supposing three continuous matrices of $W^q, W^k, W^v \in \mathbb{R}^{d\times d_k}$ (usually $d_k = d$), the calculation for the three feature sequences of Key (K), Query (Q), and Value (V) are as follows:

$$Q = C_i W^q, \quad K = C_i W^k, \quad V = C_i W^v. \tag{2}$$

The attention weights A can be calculated as follows:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right). \tag{3}$$

Finally, aforementioned attention is utilized to weight the output features as follows:

$$Z = AV. \tag{4}$$

In the third step, the weight matrix learned from the previous step is fed into an MLP layer, which consists of two fully connected layers and a nonlinear activation function between them.

Generally, ViTs require intensive computation resources and a large number of input data to mitigate the risk of overfitting, due to their lack of inductive bias [45]. Zhu et al. [46] designed a novel two-layer routing AM of BiFormer to save computation and memory resources, where each query attends to a small set of semantically relevant key–value pairs. Liu et al. [47] developed an EfficientViT by employing cascaded group attention modules to process different splits of the full feature set, minimizing computational redundancy. Fang et al. [48] proposed a multiattention joint convolution feature representation with lightweight transformer for HSI classification under small sampling rates. Zhao et al. [49] proposed a groupwise separable convolutional ViT (GSC-ViT) to effectively capture local and global features, offering a more lightweight solution. Zhang et al. [50] proposed the channel lightweight multihead self-attention and position lightweight multihead self-attention modules to reduce memory usage and computational overhead, which retains global information for each pixel or channel. While the existing approaches reduce model size by limiting ViT blocks or avoiding the quadratic complexity of self-attention, it remains a great challenge to balance the classification accuracy with parameter amounts. This article proposes a novel LSPCA module, composed of spatial pooling attention (SpaPA) and spectral pooling attention (SpePA), to replace the traditional self-attention mechanism and reduce quadratic complexity. SpaPA utilizes adaptive average and maximum pooling to aggregate smooth and salient spatial features in different directions, while SpePA uses adaptive average pooling for global spectral sequential features.

B. Multiscale Feature Extraction by the CNN for HSI Classification

CNNs are highly effective in local feature discrimination. Multiscale convolution plays a crucial role in capturing information at different granularities by extracting features at multiple levels, which enhances classification performance by improving feature utilization [51]. Xiong et al. [52] proposed an approach that adaptively focused on different spatial contexts through various convolutions with distinct kernel sizes. Han et al. [53] designed a dual-stream convolution network to learn spatial–spectral features from blocks of different scales surrounding the central pixel. Feng et al. [54] developed a multiscale CNN architecture to improve region homogeneity. Safari et al. [55] combined various CNNs to learn spatial–spectral features from multiple scales. Gong et al. [56] introduced a hybrid 2-D–3-D CNN to fuse multiscale information, enhancing the utilization of HSI data and addressing the problem of limited samples. Gao et al. [29] combined different kernel sizes in a single convolution operation to extract spatial features across multiple scales. The capsule attention network (CAN) [64] combines the activity vector with an attention-based feature extraction module and a self-weighting mechanism to improve HSI classification. HybridKAN [63] incorporates 1-D, 2-D, and 3-D KANs to enhance HSI classification performance. Although multiscale features enhance classification performance, the interaction between them often remains insufficient. Integrating these multiscale features effectively to improve feature representation and efficiency has become a crucial area of research. In this article, we introduce a shuffling technique to increase the flow and interaction of multiscale features rather than limited to single-scale features, which significantly outperforms the simpler and more direct concatenation approach.

III. Proposed Method

In this article, we propose the MSC-3DLT to enhance the accuracy and efficiency of HSI classification. The overall architecture of our method is depicted in Fig. 1. First, PCA is employed to reduce the dimensionality of HSI data by eliminating spectral redundancy. Then, the MSC module is designed to extract fine-grained features, improve information flow, and enhance spectral features. Subsequently, the 3DLT is proposed to extract smooth and salient features through the LSPCA mechanism, thereby strengthening the interaction between spatial and spectral features. Finally, a softmax classifier is used to perform HSI classification. This section focuses on the newly introduced components of MSC, 3DLT, and MALR in our method.

A. Multiscale Spatial–Spectral Shuffling Module

In HSI classification, diverse environments often cause objects to reflect complex appearances, posing great challenges in capturing comprehensive features. To address this issue, a novel MSC module is proposed, with the detailed structure shown in Fig. 2. MSC primarily comprises the multiscale 3-D feature extraction (M3DFE) module, feature grouping and shuffling (FGS) module, and spectral enhancement (SE) module.

The M3DFE performs multiple 3-D convolutions with various kernel sizes, including a special 3-D point convolution with a kernel size of $3\times1\times1$ to extract pixel-level spectral and spatial features along the spectral dimension. In addition, small- and large-scale 3-D convolutions with kernel sizes of $5\times3\times3$ and $9\times5\times5$ capture spatial and spectral neighborhood features, respectively. These convolutions effectively capture fine-grained spatial features, such as edges, contours, textures, and neighboring features across different spectral bands. Furthermore, the FGS enhances the flow of multiscale spatial–spectral features. Finally, given the importance of spectral sequence information in HSIs, the SE module emphasizes spectral information by applying spectral weights to the fused multigranular features. Unlike conventional multiscale feature extraction and fusion modules, the proposed MSC not only extracts multiscale features from each dimension but also fully integrates these multigranularity features through shuffling and spectral weighting. This leads to an HSI feature cube with rich multigranularity spatial features and strong spectral continuity for subsequent feature extraction.

Specifically, suppose that $I \in \mathbb{R}^{c\times m\times n}$ is an input HSI, where c, m, and n denote the channel numbers, length, and width, respectively. To eliminate channel redundancy in HSIs, we perform channel dimension reduction using PCA. Let the HSI after PCA dimensionality reduction be $I_{\text{PCA}} \in \mathbb{R}^{l\times m\times n}$, where

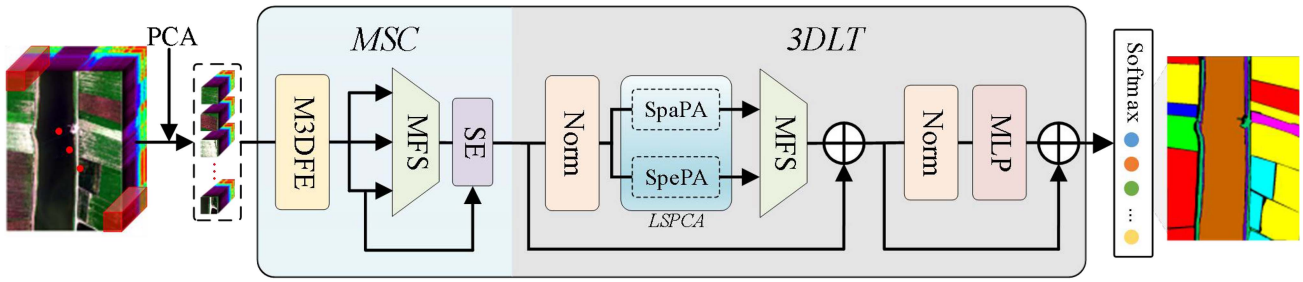


Fig. 1. Overall architecture of the proposed MSC-3DLT method for HSI classification. PCA is first applied to the input HSI for dimensionality reduction, after which the patches are fed into the MSC for multiscale feature extraction, shuffling, and enhancement by M3DFE, MFS, and SE modules, respectively. Subsequently, the enhanced features are processed through the 3DLT, where SpaPA and SpePA focus on spatial and spectral features, respectively. Finally, Softmax is performed for HSI classification.

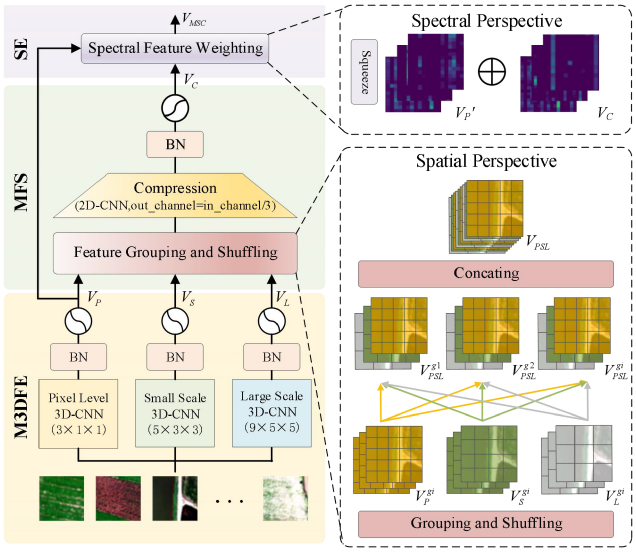


Fig. 2. Structure of the proposed MSC module, which primarily comprises three components of M3DFE, MFS, and SE. The M3DFE extracts multiscale spatial–spectral features with three different granularities, which are then fed to the MFS for feature shuffling and fusion and finally enhanced in the SE via residual connection.

l is the reduced channel dimension. In the M3DFE module, I_{PCA} is divided into multiple 3-D subblocks $V_{\text{PCA}} \in \mathbb{R}^{l\times s\times s}$ and convolved with multiscale 3-D kernels. The multiscale operation for the jth feature cube at position (x, y, z)obtained from the subblock V_{PCA} on the ith layer is obtained as follows:

$$
\begin{aligned}
vp_{ij}^{xyz} &= \sum_{d}\sum_{r=0}^{RP_{i-1}} w_{ijd}^{r} v_{(i-1)d}^{(x)(y)(z+r)} + b_{ij} \\
vs_{ij}^{xyz} &= \sum_{d}\sum_{h=0}^{HS_{i-1}}\sum_{w=0}^{WS_{i-1}}\sum_{r=0}^{RS_{i-1}} w_{ijd}^{hwr} v_{(i-1)d}^{(x+h)(y+w)(z+r)} + b_{ij} \\
vl_{ij}^{xyz} &= \sum_{d}\sum_{h=0}^{HL_{i-1}}\sum_{w=0}^{WL_{i-1}}\sum_{r=0}^{RL_{i-1}} w_{ijd}^{hwr} v_{(i-1)d}^{(x+h)(y+w)(z+r)} + b_{ij}
\end{aligned}
\tag{5}
$$

where vp, vs, and vl are the convolution features extracted under different scales, d represents the feature cube generated in intermediate convolutions, H, W, and R are the three dimensions of 3-D convolution kernels with suffixes P, S, and L indicating different scales. w and b are the parameters in weight matrix and bias, respectively. After convolution, batch normalization, and ReLU activation function, we obtain three feature cubes of $V_P \in \mathbb{R}^{l\times s\times s}$, $V_S \in \mathbb{R}^{l\times s\times s}$, and $V_L \in \mathbb{R}^{l\times s\times s}$.

In the MFS module, the aforementioned three feature cubes are divided into three groups of $V_P^{gi} \in \mathbb{R}^{g\times s\times s}$, $V_S^{gi} \in \mathbb{R}^{g\times s\times s}$, and $V_L^{gi} \in \mathbb{R}^{g\times s\times s}$ and then shuffled into feature groups to get multiple fine-grained feature matrices $V_{\text{PSL}} \in \mathbb{R}^{(3\times l)\times s\times s}$. The specific formula for FGS in MFS module is as follows:

$$
\begin{aligned}
\left[V_P^{g1}, V_P^{g2}, V_P^{\text{g3}}, \ldots, V_P^{gi}\right] &= \text{group}\left(V_P\right) \\
\left[V_S^{g1}, V_S^{g2}, V_S^{g3}, \ldots, V_S^{gi}\right] &= \text{group}\left(V_S\right) \\
\left[V_L^{g1}, V_L^{g2}, V_L^{g3}, \ldots, V_L^{gi}\right] &= \text{group}\left(V_L\right)
\end{aligned}
\tag{6}
$$

$$
V_{\text{PLS}}^{gi} = \text{Concat}\left(V_P^{gi}, V_S^{gi}, V_L^{gi}\right)
$$

$$
V_{\text{PSL}} = \text{Concat}\left(V_{\text{PSL}}^{g1}, V_{\text{PSL}}^{g2}, V_{\text{PSL}}^{g3}, \ldots, V_{\text{PSL}}^{gi}\right) \tag{7}
$$

where gi represents the group number. Subsequently, the 2-D CNN is applied to refine spatial features, which enhances local feature representation and restores the original 3-D HSI structure for weighting spectral sequences. Specifically, the jth feature map at position (x, y) in the ith layer of subblock V_{PSL} is obtained as follows:

$$
vc_{ij}^{xy} = \sum_{d}\sum_{h=0}^{H_{i-1}}\sum_{w=0}^{W_{i-1}} w_{ijd}^{hw} vc_{(i-1)d}^{(x+h)(y+w)} + b_{ij} \tag{8}
$$

where d represents the feature cube generated in intermediate convolutions, and H and W are the two dimensions of 2-D convolution kernel. Similarly, after convolution, normalization, and ReLU activation function, it will yield the output of the MFS module by the feature block V_C. To maintain the spectral sequence features and enhance the feature representation ability, the light dimension spectrum features are weighted by SE. The weights are extracted by 3-D point convolution, and the weights $V_P^{'}$ are obtained after dimension alignment, which are then weighted by residual connection. Finally, the output V_{MSC} of the MSC module is obtained as follows:

$$
V_{\text{MSC}} = V_{\text{C}} + V_P^{'}. \tag{9}
$$

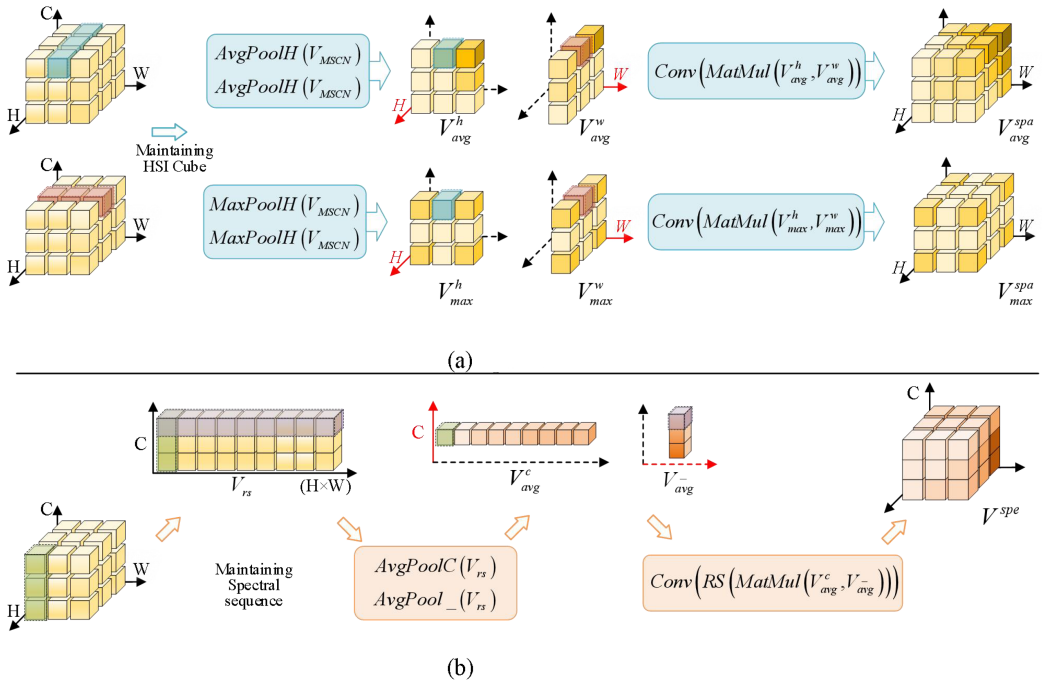


Fig. 3. Structure of the proposed LSPCA module, which mainly consists of SpaPA and SpePA. (a) SpaPA submodule. It aggregates smooth and prominent spatial features in different directions through average and maximum pooling operations, followed by interaction through matrix multiplication. (b) SpePA submodule. It aggregates global spectral sequential features via average pooling operation and then restores the feature dimension using an auxiliary matrix. (a) SpaPA Branch. (b) SpePA Branch.

B. 3-D Lightweight Transformer

Transformer excels at capturing long-range dependencies between tokens of text in the NLP field. However, when applied to HSI classification, it will inevitably damage the essential 3-D structure features since HSIs are required to be split into tokens to adapt to transformer models [57], [58]. In addition, the high dimensionality of the HSI will result in a significant increase in data volume. Furthermore, the HSI classification efficiency is also reduced for the quadratic complexity generated by multiple computations during data processing in transformer models. It is worth noting that the detailed formula is derived in Section II-A; when calculating self-attention, the complexity of linear transformation for K, Q, and V is $O(nd^2)$, the complexity of matrix multiplication in attention score QK^T is $O(n^2d)$, and that of attention weight sum AV is $O(n^2d)$.

A recent study demonstrates that the effectiveness of the transformer heavily depends on its framework structure [59], [60]. To enable transformers to effectively capture the structural features of HSIs and increase computational efficiency, we rethink the intrinsic relationship between HSI data and classification methods and propose a novel 3DLT model. The advantage of 3DLT lies in that it directly processes 3-D HSI data instead of splitting HSIs into tokens as classical transformer methods. In addition, the self-attention mechanism is substituted with a novel LSPCA mixer, which makes the spatial–spectral feature extraction efficient while significantly reducing the computational complexity, and the calculation process can be formulated as follows:

$$V_{LA} = \text{LSPCA}\left(\text{Norm}\left(V_{\text{MSC}}\right)\right) + V_{\text{MSM}} \tag{10}$$

$$V_{LT} = \text{MLP}\left(\text{Norm}\left(V_{LA}\right)\right) + V_{LA} \tag{11}$$

where Norm$(\bullet)$ represents layer normalization operation, LSPCA$(\bullet)$ denotes the lightweight spatial–spectral pooling cross-attention, which will lead to a feature cube V_{LA}. MLP$(\bullet)$ is the multilayer perceptron, and V_{LT} represents the feature cube obtained by the 3DLT module.

The proposed LSPCA mixer is the key component of the 3DLT module and comprises two branches of SpaPA and SpePA, as shown in Fig. 3(a) and (b), respectively. For the SpaPA branch, the smooth spatial features are aggregated by average pooling along spatial dimensions of H and W in the HSI, which achieves two feature cubes of $V^h_{\text{avg}} \in \mathbb{R}^{l\times s\times p}$ and $V^w_{\text{avg}} \in \mathbb{R}^{l\times p\times s}$, where p denotes the scale of adaptive pooling operation. Then, the spatial features in H and W dimensions are integrated with each other through matrix multiplication, which yields the feature cube of $V_{\text{avg}} \in \mathbb{R}^{c\times s\times s}$. The formulas are as follows:

$$\begin{aligned} V^h_{\text{avg}} &= \text{AvgPool}H\left(V_{\text{MSCN}}\right) \\ V^w_{\text{avg}} &= \text{AvgPool}W\left(V_{\text{MSCN}}\right) \\ V^{\text{spa}}_{\text{avg}} &= \text{Conv}\left(\text{MatMul}\left(V^h_{\text{avg}}, V^w_{\text{avg}}\right)\right) \end{aligned} \tag{12}$$

where AvgPool$H(\bullet)$ and AvgPool$W(\bullet)$ are the adaptive average pooling operation in H and W dimensions, respectively. MatMul$(\bullet)$ represents the matrix multiplication, and Conv$(\bullet)$ denotes the convolution operation. To enhance salient spatial features, maximum pooling is employed to extract high-frequency features. Similarly, the formulas are as follows:

$$V^h_{\text{max}} = \text{MaxPool}H\left(V_{\text{MSCN}}\right)$$

$$V_{\max}^{w} = \mathrm{MaxPool}W\left(V_{\mathrm{MSCN}}\right)$$
$$V_{\max}^{\mathrm{spa}} = \mathrm{Conv}\left(\mathrm{MatMul}\left(V_{\max}^{h}, V_{\max}^{w}\right)\right) \tag{13}$$

where MaxPool$H(\bullet)$ and MaxPool$W(\bullet)$ are adaptive maximum pooling operations in H and W dimensions. Then, the features obtained by maximum pooling and average pooling are fused as follows:

$$V^{\mathrm{spa}} = V_{\mathrm{avg}}^{\mathrm{spa}} + V_{\max}^{\mathrm{spa}}. \tag{14}$$

For the SpePA branch, the smooth spectral features are aggregated by average pooling along spectral dimension. Specifically, the input feature cube is reshaped as $V_{rs} \in \mathbb{R}^{(s\times s)\times l\times 1}$, and then, the feature aggregation in spectral dimension is performed to obtain $V_{\mathrm{Avg}}^{c} \in \mathbb{R}^{(s\times s)\times l\times p}$. To maintain the feature shape, an auxiliary feature cube $V_{\bar{\mathrm{avg}}} \in \mathbb{R}^{(s\times s)\times p\times 1}$ is constructed. After matrix multiplication, we reshape the sequential spectral features and perform convolution to obtain $V^{\mathrm{spe}} \in \mathbb{R}^{l\times s\times s}$. The formulas are as follows:

$$V_{\mathrm{avg}}^{c} = \mathrm{AvgPool}C\left(V_{rs}\right)$$
$$V_{\bar{\mathrm{avg}}} = \mathrm{AvgPool\_}\left(V_{\mathrm{reshape}}\right)$$
$$V^{\mathrm{spe}} = \mathrm{Conv}\left(\mathrm{reshape}\left(\mathrm{MatMul}\left(V_{\mathrm{avg}}^{c}, V_{\bar{\mathrm{avg}}}\right)\right)\right). \tag{15}$$

Subsequently, the spatial and spectral features are fused in a feature shuffling manner, and the formula is as follows:

$$V_{\mathrm{scf}} = \mathrm{FGS}\left(V^{\mathrm{spa}}, V^{\mathrm{spe}}\right). \tag{16}$$

Finally, softmax is adopted to perform HSI classification. In the aforementioned calculation process, the complexity of our method is $O(nd^2)$ that mainly comes from the MatMul$(\bullet)$ operation.

C. MALR Scheduler

Generally, the performance of HSI classification based on deep learning methods is influenced by the chosen training strategy. As we know, it plays a vital role in reasonably scheduling the learning rate throughout the training process. However, many researchers employ the straightforward approach of a fixed learning rate (FLR) to train models, which, while requiring no additional parameters, has inherent limitations. The FLR may slow the initial training phase and cause instability as the model approaches its optimal solution in later epochs. To address these drawbacks, researchers have adopted a linearly decreasing learning rate (LDLR) strategy. This approach begins with a higher learning rate to expedite early training and gradually reduce it after a fixed number of epochs, stabilizing the model as it approaches convergence. While this method improves upon the FLR by adjusting the learning rate periodically, it is not without its shortcomings. The decision to reduce the learning rate is predetermined, independent of the model's real-time performance. Consequently, the static nature of the LDLR often necessitates extensive trial-and-error experiments to identify an appropriate interval for learning rate reduction. This inflexibility limits its effectiveness in dynamic parameter scheduling.

Algorithm 1: MALR Scheduler Algorithm.

Input:
Initial learning rate lr_0; attenuation factor f_a; patience period p_p; cooling down period p_c; schedulers for loss and accuracy scheduler$_{\mathrm{loss}}$ and scheduler$_{\mathrm{acc}}$; minimum learning rate $lr_{\min}$; maximum change threshold $T_{\max}$.
Output:
New schedulers of scheduler$_{\mathrm{loss}}$ or scheduler$_{\mathrm{acc}}$.
1: Initialization: epoch for loss changes of $\mathrm{Epoch}_{\mathrm{Loss}} = 0$; epoch for verification accuracy changes of $\mathrm{Epoch}_{\mathrm{Acc}} = 0$.
2: **for** i $= 1$ to $\triangle$ **do**
3: Obtain the validation accuracy Acc_i and validation loss $\mathrm{Loss}_{\mathrm{i}}$ for each epoch during training process.
4: **if** $\mathrm{Loss}_i - \mathrm{Loss}_{i-1} \neq 0\&\mathrm{Epoch}_{\mathrm{loss}} > P_c$ **:** $\mathrm{Epoch}_{\mathrm{Loss}} += 1$
else $\mathrm{Epoch}_{\mathrm{Loss}} = 0$.
5: **if** $\mathrm{Acc}_i - \mathrm{Acc}_{i-1} \neq 0\&\mathrm{Epoch}_{\mathrm{acc}} > P_c$**:** $Epoch_{acc} += 1$
else $\mathrm{Epoch}_{\mathrm{acc}} = 0$.
6: **if** $|lr_i - lr_0| \leq T_{\max}\& lr_i \geq lr_{\min}$**:** $lr_i = f_a \times lr_{i-1}$
Update scheduler$_{\mathrm{loss}}$ **or** scheduler$_{\mathrm{acc}}$
7: Reset $\mathrm{Epoch}_{\mathrm{loss}} = 0$ and $\mathrm{Epoch}_{\mathrm{Acc}} = 0$.
8: **end for**

To overcome these challenges, we design a novel training strategy, i.e., MALR scheduler. Designed to enhance the flexibility and efficiency of our proposed MSC-3DLT method, the MALR adapts the learning rate in real time, informed by feedback from multiple metrics during the training process. Unlike the FLR and the LDLR, the MALR scheduler dynamically adjusts the learning rate based on the model's fitness at each epoch, responding to changes in gradients and obviating the need for manual hyperparameter tuning. The adaptive nature of the MALR offers several advantages. By tailoring the learning rate to the model's performance, the MALR minimizes the risk of overfitting and enhances generalization. This adaptive adjustment accelerates convergence, enabling the model to reach optimal performance in fewer epochs. In the MALR, we stipulate that if multimetrics (loss value or validation accuracy) do not improve in a certain epoch and exceed the cooling period, the weight will be used to decay the learning rate. In addition, we specify the minimum learning rate and the maximum change threshold to control the bounds of the learning rate. The specific algorithm of the MALR is shown in Algorithm 1.

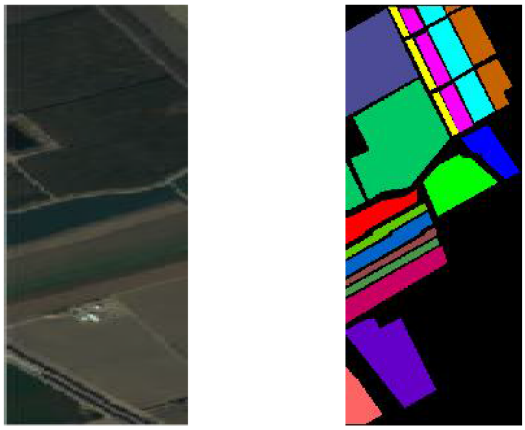
IV. Experiments and Analysis

A. Data Description

1) *Salinas Valley (SV)* was captured by the airborne AVIRIS sensor over SV agricultural area. The image has a spatial dimension of 512 × 217 pixels, with 204 bands available for experiments. This dataset contains 54 129 labeled samples, encompassing 16 classes of land cover. The details of these classes and sampling rates are shown in Table I,

TABLE I
TRAINING, VALIDATION, AND TESTING SAMPLES IN THE SV DATASET

No.	Land Cover Class	Train	Val	Test
1	Brocoli green weeds 1	11	11	1987
2	Brocoli green weeds 2	19	19	3688
3	Fallow	10	10	1956
4	Fallow rough plow	7	7	1380
5	Fallow smooth	14	14	260
6	Stubble	20	20	3919
7	Celery	18	18	3543
8	Grapes untrained	57	57	11157
9	Soil vinyard develop	32	32	6139
10	Corn senesced green weeds	17	17	3244
11	Lettuce romaine 4wk	6	6	1056
12	Lettuce romaine 5wk	10	10	1907
13	Lettuce romaine 6wk	5	5	906
14	Lettuce romaine 7wk	6	6	1058
15	Vinyard untrained	37	37	7194
16	Vinyard vertical trellis	10	10	1787
Total		279	279	51181



(a) (b)

Fig. 4. SV dataset. (a) False-color map. (b) Ground truth map.

while the ground truth map and sample distribution are shown in Fig. 4.

2) *WHU-Hi-LongKou (LK)* was collected by the Headwall Nano-Hyperspec sensor in 2018 over Longkou Town in Hubei Province, China. The image is with a spatial dimension of 550 × 400 pixels, containing 270 bands for experiments. It contains 204 542 labeled samples, including nine types of land cover. The details of these types and sampling rates are shown in Table II, while the ground truth map and sample distribution are shown in Fig. 5.
3) *XuZhou (XZ)* was captured by the airborne HySpex hyperspectral camera over the suburban area of Xuzhou. The image has 500 × 260 pixels with 436 bands available for experiments. It contains 68 882 labeled samples and nine types of land cover. The details of these types and sampling rates are shown in Table III, while the ground truth map and sample distribution are shown in Fig. 6.
4) *DFC2018 Houston (HS2018)* was issued by the University of Houston and utilized in the 2018 IEEE GRSS Data Fusion competition. The image has 601 × 2384 pixels and contains 50 bands available for experiments. It includes 502 856 labeled samples and 20 land cover types. The details of these types and sampling rates are shown in Table IV, while the ground truth map and sample distribution are shown in Fig. 7.

TABLE II
TRAINING, VALIDATION, AND TESTING SAMPLES IN THE LK DATASET

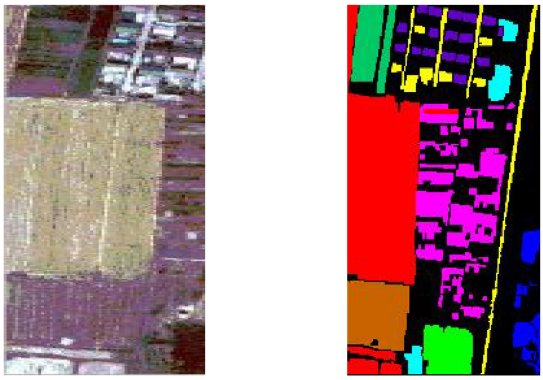
No.	Land Cover Class	Train	Val	Test
1	Corn	70	70	34371
2	Cotton	17	17	8340
3	Sesame	7	7	3017
4	Broad-leaf soybean	127	127	62958
5	Narrow-leaf soybean	9	9	4133
6	Rice	24	24	11806
7	Water	135	135	66786
8	Roads and houses	15	15	7094
9	Mixed weed	11	11	5207
Total		415	415	203712

(a) (b)

Fig. 5. LK dataset. (a) False-color map. (b) Ground truth map.

TABLE III
TRAINING, VALIDATION, AND TESTING SAMPLES IN THE XZ DATASET

No.	Land Cover Class	Train	Val	Test
1	Bareland 1	132	132	26135
2	Lakes	21	21	3985
3	Coals	14	14	2755
4	Cement	27	27	5161
5	Crops 1	66	66	13052
6	Trees	13	13	2410
7	Bareland 2	35	35	6921
8	Crops 2	24	24	4729
9	Red Tiles	16	16	3038
Total		348	348	68186

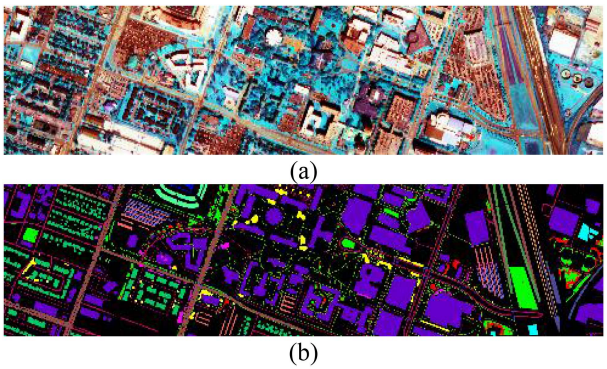


(a) (b)

Fig. 6. XZ dataset. (a) False-color map. (b) Ground truth map.

TABLE IV
TRAINING, VALIDATION, AND TESTING SAMPLES IN THE HS2018 DATASET

No.	Land Cover Class	Train	Val	Test
1	Healthy grass	98	98	9603
2	Stressed grass	326	326	31850
3	Artificial turf	7	7	670
4	Evergreen trees	136	136	13323
5	Deciduous trees	51	51	4919
6	Bare earth	46	46	4424
7	Water	3	3	260
8	Residential building	398	398	38976
9	Non-residential building	2238	2238	219276
10	Road	459	459	44948
11	Sidewalks	341	341	33347
12	Crosswalks	16	16	1486
13	Major thoroughfares	464	464	45420
14	Highways	99	99	9667
15	Railways	70	70	4797
16	Paved parking lots	115	115	11270
17	Unpaved parking lots	2	2	142
18	Cars	66	66	6415
19	Trains	54	54	5261
20	Stadium seats	69	69	6686
Total		5058	5058	492740



(a)

(b)

Fig. 7. HS2018 dataset. (a) False-color map. (b) Ground truth map.

B. Evaluation Criteria and Experimental Settings

1) Evaluation Criteria: To quantitatively evaluate the performance of our method, we employ the evaluation metrics of overall accuracy (OA), average accuracy (AA), Kappa coefficient (K), and the classification accuracy of each individual land cover type. These metrics are utilized to measure the classification accuracy and consistency of each method, with higher values being better performance. In addition, the metrics of parameter numbers (Params) and FLOPs are adopted to measure model complexity and computational overhead, with smaller values being preferable.

2) Environment Configuration and Parameter Settings: The proposed method is implemented using PyTorch 1.11.0 on an Ubuntu 20.04.4 operating system, with experiments conducted on a workstation equipped with a Platinum 8352V CPU, an RTX 4090 (24 GB) GPU, and 90-GB RAM. In our experiments, the Adam optimizer is utilized, batch size is set to 128, patch size is set to 11, and training epochs are set to only 50 in total. The initial learning rate is set to 1e-2 to accelerate the training speed in early epochs. We adopt the proposed training strategy of the MALR scheduler to accurately train our method in later epochs. Among them, the parameters of attenuation factor, patience level, cooling down period, maximum change threshold, and minimum learning rate are set to 0.5, 7 (epochs), 2 (epochs), 1e-3, and 1e-5, respectively. To guarantee the fairness of comparisons, we repeat each experiment ten times and then take their mean and standard deviations as the final results.

C. Ablation Experiments

To evaluate the influence of each component in the proposed MSC-3DLT method on classification performance, ablation experiments are conducted on the SV dataset. We randomly select 0.5% samples as the training and validation sets, while the remaining as the test set. The classical transformer is adopted as the baseline method to verify the two newly proposed components of MSC and 3DLT:

1) NET-1: classical transformer method;
2) NET-2: proposed 3DLT model;
3) NET-3: MSC + 3DLT (without FGS);
4) NET-4: MSC + 3DLT (without SpaPA);
5) NET-5: MSC + 3DLT (without SpePA);
6) NET-6: MSC + 3DLT.

The details of aforementioned six network combinations and ablation experimental results are reported in Table V. From the HSI classification results of NET-1, it is evident that a single transformer block leads to very poor classification performance for few training samples.

1) Effectiveness of the 3DLT by Comparing NET-1 With NET-2: NET-1 observes that the 3DLT greatly improves HSI classification performance in terms of all metrics obtained by NET-2. The reason mainly lies in that the 3DLT extracts the essential spatial–spectral structure features from the HSI since it directly processes 3-D HSI data without flattening operations as classical transformers. In addition, NET-2 significantly reduces the Params and FLOPs of the model, which uses LSPCA to extract spatial and spectral features without destroying the 3-D

TABLE V
DETAILS OF EACH NETWORK AND ABLATION EXPERIMENTAL RESULTS

	Transformer	MSC	(FGS)	3DLT	(SpaPA)	(SpePA)	OA(%)	AA(%)	K×100	Params(k)	FLOPs(M)
NET-1	✓						82.24±5.97	82.24±5.97	82.24±5.97	42.264	25.24
NET-2				✓	✓	✓	97.57±0.50	97.24±0.72	97.30±0.55	**10.456**	**1.22**
NET-3		✓		✓	✓	✓	97.82±0.36	97.65±0.64	97.57±0.40	26.362	3.94
NET-4		✓	✓	✓	✓		97.64±0.46	97.31±0.67	97.37±0.51	26.074	3.90
NET-5		✓	✓	✓		✓	97.77±0.47	97.35±0.59	97.52±0.52	26.074	3.88
NET-6		✓	✓	✓	✓	✓	**98.09±0.40**	**97.98±0.57**	**97.88±0.45**	26.362	3.94

* FGS is a component of MSC while SpaPA and SpePA are components of 3DLT
The bold value indicates the optimal value in each evaluation metric.

HSI structure. The advantage of the 3DLT is that it effectively aggregates smooth features and emphasizes salient features while reducing quadratic complexity.

2) Effectiveness of MSC by Comparing NET-2 With NET-6: As observed, NET-6 increases OA, AA, and Kappa coefficient by 0.52%, 0.74%, and 0.58, respectively, which is mainly attributed to the inclusion of MSC that extracts features from original HSI using multiscale 3-D convolution. This module effectively captures local edges and textures and maintains spectral continuity. These multiple fine-grained features are fused, and the spectral features are weighted to provide a richer 3-D data cube for subsequent 3DLT feature aggregation.

3) Effectiveness of FGS by Comparing NET-3 With NET-6: NET-3 utilizes the traditional feature concatenation method, while NET-6 employs the proposed FGS method to enrich the granularities of spatial and spectral features. The advantage of FGS lies in that it can shuffle different groups of fine-grained features by reconstructing regions with higher correlation close to each other. This method enhances the representation of spatial and spectral features by fusing the detailed differences of local features and weights the spectral features through residual connection to maintain the continuity of spectral features.

4) Effectiveness of SpaPA and SpePA by Comparing NET-4 and NET-5 With NET-6: It is noted that the former two methods retain either SpaPA or SpePA module. By comparison, the classification results clearly demonstrate that simultaneously utilizing both SpaPA and SpePA further improves classification accuracy by effectively complementing features between each other.

On the whole, NET-6, i.e., the proposed MSC-3DLT method, outperforms all other counterpart methods in terms of OA, AA, and Kappa coefficient with few fluctuations of 0.40%, 0.57%, and 0.45, respectively

D. Experimental Results and Comparative Analysis

To quantitatively and qualitatively validate the effectiveness of the proposed MSC-3DLT method in HSI classification with small sampling rates, we compare our method with 11 SOTA methods of HybridSN [27], RSSAN [32], LSSCM [61], SPRN [62], GAHT [37], SSFTT [39], MorphFormer [40], HybridKAN [63], CAN [64], GSC-ViT[48], and DBCTNet [65] on SV, LK, XZ, and HS2018 datasets. The optimal values are highlighted in bold, while the suboptimal values underlined to make the quantitative comparisons clearer.

The classification results of various methods on the SV, LK, XZ, and HS2018 datasets, with small sampling rates of 0.5%, 0.2%, 0.5%, and 1%, respectively, are presented in Tables VI–IX, respectively. For the SV dataset, our method outperforms all other methods, achieving the highest classification accuracy in terms of OA, AA, and Kappa coefficient, while also demonstrating excellent performance in Params and FLOPs.

Specifically, HybridSN ranks second in OA, but its performance falls short of our MSC-3DLT by only 0.14%. HybridSN extracts spatial–spectral features and enhances spatial features through multiple layers of 2-D and 3-D convolutions, which, while improving classification accuracy, leads to significant increases in Params and FLOPs. LSSCM, on the other hand, reduces Params and FLOPs by employing deep separable convolution, but it focuses too much on spatial information at the expense of spectral sequential features, resulting in a decline in classification accuracy. RSSAN and SPRN rely on AMs to identify effective features and reduce Params, but their overemphasis on local features leads to poor global feature extraction and, thus, diminished classification accuracy. HybridKAN achieves the smallest FLOPs but at the cost of classification performance. The CAN, while offering relatively good classification accuracy, has Params that reach 429.217k, far exceeding those of our method. The proposed MSC-3DLT method, by contrast, combines 2-D and 3-D convolutions and enhances local spatial features through the MSC module, all while preserving spectral feature continuity and avoiding the stacking design commonly employed in other methods.

When compared to transformer-based methods, the proposed MSC-3DLT shows significant improvements. It avoids the structural degradation commonly seen in other transformer methods when HSIs are split into tokens. In addition, the self-attention mechanism in typical transformer models not only fails to capture long-range dependencies but also contributes to an increase in Params and FLOPs, particularly at low sampling rates due to its quadratic complexity. For example, the GAHT method uses group convolutions and transformer blocks to interact local and global features, resulting in hefty 972.624k parameters. In contrast, SSFTT and MorphFormer use fewer parameters of 148.488k and 202.8k by adopting a single transformer block and morphology-based self-attention, respectively, but they still struggle with performance under small sampling rates. The GSC-ViT, with just 104.976k Params, suffers from poor classification accuracy in these conditions. DBCTNet, which combines 3-D CNN and ConvTE branches for extracting spatial and

TABLE VI
CLASSIFICATION ACCURACY AND MODEL COMPLEXITY OF THE CNN-BASED METHOD AND THE TRANSFORMER-BASED METHOD ON THE SV DATASET

	HybridSN	RSSAN	LSSCM	SPRN	GAHT	SSFTT	MorphFormer	HybridKAN	CAN	GSC-ViT	DBCTNet	MSC-3DLT
1	**99.98±0.06**	96.68±2.97	96.84±7.26	99.81±0.49	99.80±0.59	99.44±1.42	99.83±0.29	76.74±0.25	99.54±0.82	97.03±7.10	97.02±2.51	99.81±0.40
2	99.07±0.89	98.45±1.72	99.44±1.03	**100.00±0.00**	99.84±0.20	99.38±1.37	99.69±0.69	92.93±0.11	99.89±0.22	99.64±0.66	99.50±0.66	98.40±1.54
3	99.53±0.46	88.90±7.24	79.38±7.44	92.10±10.88	95.62±4.28	**99.73±0.65**	87.67±7.51	79.92±5.96	98.71±1.66	81.92±19.10	95.84±2.63	99.40±0.93
4	98.62±1.72	97.14±3.85	98.36±1.50	98.54±1.48	**99.32±0.88**	98.75±0.99	97.70±3.08	91.09±2.42	98.15±1.87	97.78±3.38	98.71±1.25	98.91±0.65
5	95.04±2.21	94.24±5.52	95.08±3.85	95.91±2.62	95.71±3.13	**97.69±1.78**	93.44±3.24	95.48±3.11	97.26±3.52	94.18±4.77	96.11±1.28	95.04±1.53
6	99.81±0.23	97.48±1.94	99.33±0.73	99.87±0.24	99.91±0.17	99.32±1.32	99.23±1.29	**99.9±0.18**	95.67±3.80	99.58±0.84	99.15±1.32	99.06±1.91
7	99.96±0.05	97.74±2.88	99.39±0.71	99.94±0.10	99.90±0.11	99.46±0.80	99.46±0.56	90.51±3.31	99.84±0.21	99.55±1.35	97.61±2.26	**99.99±0.03**
8	**97.64±0.88**	84.33±6.50	83.22±4.82	88.83±3.15	89.41±5.33	93.74±1.96	90.14±3.64	76.11±8.83	94.34±4.17	89.34±4.28	90.69±2.51	97.41±0.93
9	99.85±0.20	98.90±0.85	99.72±0.61	99.93±0.11	99.86±0.31	99.97±0.06	99.33±0.64	99.62±0.24	**99.98±0.05**	98.90±3.14	98.88±0.75	99.65±0.46
10	95.21±1.86	90.69±5.49	89.09±5.27	93.29±3.16	95.59±1.73	95.93±2.33	93.93±4.08	78.18±5.54	**96.45±2.35**	93.97±3.57	93.82±0.91	95.54±1.64
11	98.21±1.49	85.14±10.84	41.38±32.82	92.93±7.36	93.26±5.43	96.05±4.66	94.89±3.20	57.66±0.66	97.92±2.05	90.33±12.40	90.15±11.14	**98.75±2.13**
12	98.90±2.03	97.83±3.07	95.23±9.04	**99.98±0.02**	98.80±2.81	98.49±2.99	97.69±4.47	75.04±6.31	91.54±3.88	97.23±5.08	99.59±0.54	99.30±0.60
13	89.83±9.24	97.48±3.95	93.22±5.75	**98.57±1.01**	97.36±4.53	95.14±6.67	97.36±5.00	77.35±2.24	78.48±14.92	98.18±3.69	92.19±7.41	92.64±6.38
14	96.94±1.92	96.20±3.42	96.57±3.22	98.42±1.06	**99.37±0.85**	96.47±2.64	97.83±1.63	93.58±3.65	93.11±4.89	96.90±2.81	98.09±1.02	97.36±2.52
15	96.13±2.22	78.02±11.67	77.80±11.78	92.25±1.89	88.35±4.77	91.17±2.78	90.26±5.02	65.47±5.22	96.05±2.97	86.76±6.54	79.95±5.66	**97.74±0.67**
16	98.62±0.25	93.92±2.60	89.42±6.84	97.34±1.75	97.01±2.29	98.07±0.92	96.00±3.14	97.89±2.91	**99.04±0.82**	96.12±2.98	94.88±3.86	98.65±0.33
OA%	97.95±0.46	91.08±1.52	89.69±2.31	95.39±0.82	95.19±0.72	96.61±0.43	94.84±0.53	82.78±5.05	96.62±1.01	93.80±0.72	93.57±0.60	**98.09±0.40**
AA%	97.71±0.64	93.32±1.08	89.59±2.78	96.73±1.05	96.82±0.71	97.43±0.66	95.90±0.59	83.61±7.11	96.00±1.03	94.84±1.47	95.14±0.83	**97.98±0.57**
K×100	97.72±0.51	90.07±1.69	88.50±2.59	94.87±0.91	94.65±0.80	96.22±0.48	94.25±0.59	80.84±5.67	96.24±1.12	93.10±0.80	92.84±0.67	**97.88±0.45**
Params/K	5122.176	108.432	**12.64**	183.348	972.624	148.488	202.8	134.512	429.217	104.976	30.888	26.362
FLOPs/M	247.68	23.53	1.49	9.04	47.61	11.40	36.39	**0.316**	1.488	6.62	11.9	3.94

The bold value indicates the optimal value while the underlined value for the suboptimal value in each evaluation metric.

TABLE VII
CLASSIFICATION ACCURACY AND MODEL COMPLEXITY OF THE CNN-BASED METHOD AND THE TRANSFORMER-BASED METHOD ON THE LK DATASET

NO.	HybridSN	RSSAN	LSSCM	SPRN	GAHT	SSFTT	MorphFormer	HybridKAN	CAN	GSC-ViT	DBCTNet	MSC-3DLT
1	99.73±0.17	99.43±0.43	99.28±0.72	99.73±0.37	99.43±0.44	**99.77±0.22**	99.51±0.26	93.25±1.28	99.24±0.10	99.48±0.67	99.67±0.18	99.63±0.13
2	97.09±1.90	88.92±4.71	91.79±6.21	90.69±7.71	94.94±3.92	97.34±3.10	97.22±1.67	52.65±21.62	88.40±3.58	93.65±5.76	93.26±6.6	**97.65±1.51**
3	99.08±0.37	89.35±5.68	83.02±12.35	90.31±3.98	92.86±4.87	93.96±2.29	90.64±6.36	57.1±29.21	79.19±3.14	93.74±3.50	**99.44±0.78**	98.71±0.52
4	98.12±0.45	98.28±0.74	98.95±0.76	98.87±0.54	98.44±0.37	98.02±0.72	99.06±0.47	90.05±5.21	98.60±0.53	99.05±0.52	**99.20±0.34**	98.48±0.44
5	91.43±4.06	81.62±7.00	62.69±26.29	90.05±4.21	92.56±3.96	94.35±3.05	87.10±5.44	85.62±4.05	65.60±11.43	87.11±6.30	86.56±9.97	**95.00±3.01**
6	99.21±0.67	97.92±1.94	97.59±2.00	99.16±1.16	97.16±4.20	99.45±0.52	97.26±1.93	96.05±2.81	99.29±0.68	98.56±1.35	**99.75±0.20**	99.46±0.39
7	99.49±0.30	99.76±0.37	99.95±0.05	99.88±0.06	99.33±0.67	99.83±0.15	99.92±0.08	99.42±0.39	**99.96±0.03**	99.80±0.26	99.94±0.07	99.66±0.10
8	92.56±2.37	91.94±3.02	91.34±4.13	92.39±6.05	89.09±7.92	94.10±2.16	90.14±4.26	85.35±6.33	94.10±2.43	90.51±5.75	91.77±3.63	**94.34±2.26**
9	81.53±3.35	81.22±8.82	46.66±17.86	88.10±5.34	81.54±4.80	89.62±4.63	87.21±6.29	73.41±8.56	**90.32±2.60**	82.08±5.53	80.65±9.23	80.95±3.62
OA%	98.13±0.38	97.43±0.38	96.39±0.91	98.22±0.31	97.72±0.44	96.61±0.43	94.84±0.53	90.95±1.36	97.45±0.16	98.07±0.40	98.11±0.59	**98.42±0.18**
AA%	95.36±0.90	92.05±1.42	85.70±4.78	94.35±1.22	93.93±1.02	**97.43±0.66**	95.90±0.59	74.6±5.15	90.52±1.38	93.78±1.06	94.47±1.85	95.99±0.61
K×100	97.54±0.50	96.62±0.49	95.23±1.20	97.66±0.41	97.01±0.57	96.22±0.48	94.25±0.59	88.08±6.24	96.64±0.21	97.46±0.52	97.53±0.77	**97.93±0.24**
Params/K	5122.176	108.432	**13.24**	183.348	972.624	148.488	202.8	132.489	487.001	104.976	40.265	26.187
FLOPs/M	247.68	28.21	1.62	9.16	74.15	11.40	48.21	**0.316**	2.016	10.25	15.84	3.94

The bold value indicates the optimal value while the underlined value for the suboptimal value in each evaluation metric.

TABLE VIII
CLASSIFICATION ACCURACY AND MODEL COMPLEXITY OF THE CNN-BASED METHOD AND THE TRANSFORMER-BASED METHOD ON THE XZ DATASET

NO.	HybridSN	RSSAN	LSSCM	SPRN	GAHT	SSFTT	MorphFormer	HybridKAN	CAN	GSC-ViT	DBCTNet	MSC-3DLT
1	96.50±1.54	96.84±0.82	96.27±0.81	97.57±0.39	97.38±0.61	96.95±0.90	97.62±0.89	92.44±1.89	**98.16±0.76**	97.71±0.84	97.90±0.44	97.92±0.67
2	76.05±38.16	95.70±2.51	98.72±0.56	98.95±0.31	**99.11±0.35**	97.74±1.81	97.66±2.22	90.52±3.70	98.62±0.31	97.70±1.55	98.86±0.25	99.05±0.43
3	48.72±41.71	81.77±9.95	83.88±9.26	80.52±5.75	78.69±6.28	86.36±9.53	**90.64±6.91**	60.8±8.65	86.73±10.20	89.30±6.74	81.59±6.20	83.06±5.89
4	71.87±31.44	92.67±4.39	86.94±4.67	95.95±1.68	95.30±1.63	95.94±1.95	93.79±4.24	76.46±12.14	93.71±2.81	94.31±3.59	**96.23±1.10**	95.43±1.38
5	94.14±5.62	97.04±1.62	96.94±1.48	98.17±1.01	**98.23±0.58**	98.04±0.90	97.46±1.34	89.26±2.21	96.72±1.64	98.01±0.88	98.01±0.54	98.06±0.58
6	54.54±39.47	86.07±5.52	77.25±10.72	98.87±1.59	98.45±2.44	96.09±2.14	91.86±6.23	86.61±5.65	95.83±2.88	93.40±3.42	97.53±2.34	**98.91±0.70**
7	96.53±3.12	96.35±2.13	97.78±1.38	98.02±0.92	**98.03±1.15**	97.92±0.82	97.46±0.86	84.8±3.36	97.39±1.31	97.57±1.56	96.92±1.33	97.71±1.22
8	68.81±34.96	96.95±1.94	98.12±1.40	99.25±0.84	**99.51±0.43**	98.21±0.98	98.74±1.22	74.56±5.21	99.01±0.60	98.29±0.93	99.03±1.35	99.04±1.44
9	53.67±41.21	96.32±2.07	93.60±5.00	98.54±0.93	**98.95±0.82**	98.45±2.08	93.98±3.57	88.94±11.62	97.98±1.06	98.05±0.88	98.38±1.44	98.33±1.13
OA%	85.75±12.44	95.44±0.86	94.83±0.53	97.20±0.43	97.05±0.36	96.92±0.58	96.72±0.98	86.54±4.87	97.00±0.12	97.06±0.37	97.18±0.25	**97.33±0.33**
AA%	73.43±23.82	93.30±1.67	92.17±1.14	96.20±0.72	95.96±0.77	96.19±1.09	95.47±1.58	81.45±5.78	96.02±0.53	96.04±0.47	96.05±0.77	**96.39±0.60**
K×100	81.66±16.20	94.22±1.09	93.43±0.67	96.45±0.54	96.26±0.45	96.10±0.73	95.83±1.24	82.89±6.14	96.19±0.16	96.27±0.46	96.42±0.32	**96.62±0.42**
Params/K	5121.273	211.823	**15.90**	190.992	966.281	148.033	387.945	132.489	564.185	163.913	64.169	20.241
FLOPs/M	247.68	40.04	1.94	9.46	47.33	11.40	77.93	**0.316**	2.726	10.42	25.72	3.12

The bold value indicates the optimal value while the underlined value for the suboptimal value in each evaluation metric.

TABLE IX
CLASSIFICATION ACCURACY AND MODEL COMPLEXITY OF THE CNN-BASED METHOD AND THE TRANSFORMER-BASED METHOD ON THE HS2018 DATASET

	HybridSN	RSSAN	LSSCM	SPRN	GAHT	SSFTT	MorphFormer	HybridKAN	CAN	GSC-ViT	DBCTNet	MSC-3DLT
1	67.15±7.05	72.68±13.00	80.49±7.50	77.82±6.48	82.29±2.23	73.49±5.41	82.04±6.31	58.66±6.16	71.91±6.98	77.45±7.50	**95.05±0.25**	78.16±4.98
2	89.55±1.31	87.01±6.18	90.58±3.04	92.01±5.02	94.11±0.93	94.46±0.95	92.46±2.89	85.31±4.51	92.02±1.69	94.90±1.56	**99.85±1.02**	89.75±2.56
3	50.03±37.05	69.21±17.13	80.80±16.53	**99.31±0.78**	95.67±3.85	98.12±3.47	88.16±6.93	24.24±24.12	95.64±4.00	97.28±2.95	89.05±0.56	96.85±6.80
4	82.57±2.71	89.99±3.40	94.48±0.81	93.21±2.58	94.94±1.62	94.79±2.32	**95.14±1.27**	74.05±12.88	92.85±1.53	93.67±1.74	66.17±6.31	93.35±2.04
5	55.27±10.98	42.99±8.19	63.73±5.34	69.46±11.69	80.82±5.19	84.13±2.00	71.73±1.05	44.24±17.76	85.04±2.71	77.71±2.74	**95.28±2.11**	86.70±3.01
6	94.16±3.53	61.58±12.31	89.23±3.79	90.43±7.12	94.78±2.89	96.71±1.65	93.25±5.05	81.41±19.44	95.86±2.62	92.73±7.40	**99.59±1.57**	98.47±2.29
7	15.77±3.03	61.76±19.32	78.70±17.29	74.94±17.52	59.92±13.08	72.85±20.93	22.44±15.51	7.48±9.43	82.67±15.17	**89.85±11.14**	80.53±4.97	88.73±11.71
8	94.31±2.52	72.08±5.17	83.28±2.68	85.90±4.11	89.70±1.84	91.75±0.45	90.03±0.68	89.40±3.51	90.37±1.60	89.48±2.63	**97.84±0.31**	94.88±1.94
9	**98.04±0.64**	91.23±2.01	95.36±1.03	97.11±0.80	97.32±0.57	97.49±0.62	97.25±0.52	97.66±0.78	97.09±0.64	97.13±0.48	76.05±3.96	97.43±0.46
10	73.42±5.63	50.82±7.14	61.93±4.80	66.36±5.13	75.93±2.41	76.47±1.61	69.57±3.99	69.68±3.36	72.52±2.71	73.91±2.23	59.35±9.11	**78.22±2.92**
11	60.65±2.93	37.95±2.97	51.77±2.48	62.00±2.79	63.91±2.95	**69.10±1.74**	61.93±3.19	48.86±7.41	66.78±3.71	66.94±2.86	18.17±6.33	66.38±4.21
12	5.65±4.67	1.97±0.45	4.06±1.38	15.36±5.16	13.45±3.70	15.07±5.40	4.87±3.63	3.98±3.23	13.55±2.30	11.52±1.75	**82.68±2.09**	20.16±3.98
13	**88.26±3.03**	60.17±3.73	73.44±3.78	83.00±7.50	85.04±0.77	85.40±2.97	87.16±3.26	84.06±3.56	84.36±2.20	86.13±3.38	75.08±3.68	88.11±1.88
14	92.40±3.51	63.62±4.80	83.22±3.69	85.23±2.81	91.76±2.08	86.98±3.03	88.73±5.09	85.99±7.43	86.89±5.42	**93.73±1.34**	93.67±5.99	92.17±2.43
15	95.43±4.54	82.04±4.41	94.74±2.82	98.34±0.87	98.19±1.91	**99.22±0.19**	96.76±1.48	89.91±8.57	98.99±0.83	98.41±1.06	88.49±4.33	97.99±0.77
16	87.51±3.36	46.90±9.66	86.31±4.65	86.16±0.41	92.88±1.62	92.35±2.32	91.40±3.18	87.58±3.33	92.65±1.54	91.59±1.94	69.76±5.22	**93.45±1.68**
17	32.39±32.57	43.94±17.40	57.18±23.94	**96.71±4.65**	68.73±15.75	80.14±8.68	16.62±14.12	16.47±15.97	65.50±21.92	64.93±32.43	65.23±5.16	77.54±18.59
18	87.38±5.90	21.20±6.55	60.77±11.51	82.66±1.80	86.01±4.27	86.19±2.35	86.38±4.14	79.03±19.53	81.70±4.95	88.16±7.84	**94.02±3.55**	87.63±2.50
19	87.76±5.03	51.28±12.47	87.64±4.40	76.19±8.93	94.75±2.15	91.77±2.76	94.14±2.38	79.85±12.99	90.03±2.70	**96.60±3.60**	90.66±2.84	91.67±3.50
20	98.63±0.78	71.45±8.04	93.58±2.49	97.30±0.58	96.67±1.83	97.79±0.88	94.91±3.89	93.22±6.94	98.29±1.45	96.53±2.48	95.01±2.13	**98.96±0.91**
OA%	89.04±1.56	74.60±0.99	84.11±0.42	87.50±0.57	89.99±0.26	90.47±0.24	89.04±0.13	85.56±3.11	89.25±0.33	89.95±0.19	87.04±2.14	**90.86±0.30**
AA%	72.82±4.86	58.99±3.13	75.57±3.73	81.48±1.80	82.84±1.53	84.21±1.01	76.25±1.05	65.05±7.85	82.74±1.95	83.93±2.12	80.87±3.56	**85.83±1.36**
K×100	85.70±2.03	66.70±1.08	79.26±0.55	83.65±0.73	86.95±0.32	87.58±0.30	85.71±0.16	81.14±4.02	86.01±0.41	86.91±0.26	83.25±3.99	**88.13±0.37**
Params/K	5122.69	54.63	10.44	178.17	745.81	148.75	79.21	134.512	345.877	79.12	**8.775**	54.8
FLOPs/M	247.68	12.65	1.2	8.76	36.49	11.4	8.81	**0.316**	0.733	5.00	2.5871	7.74

The bold value indicates the optimal value while the underlined value for the suboptimal value in each evaluation metric.

spectral features, optimizes Params and FLOPs using convolutional spectral projection and multihead self-attention. The proposed MSC-3DLT, however, processes the 3-D structure of HSI data to maintain feature integrity, effectively reducing Params and FLOPs through the LSPCA mechanism, and excels in classification at small sampling rates. Overall, the proposed MSC-3DLT method achieves highly competitive performance on the SV dataset, considering both low spatial and spectral resolutions.

On the LK dataset, with a sampling rate of 0.2%, the high spatial resolution of 0.436 m allows CNN-based methods, such as HybridSN, RSSAN, LSSCM, and SPRN, to achieve OA values around 98%, thanks to their focus on spatial features. As the number of samples and spatial resolutions increase, transformer-based methods improve the classification performance. For example, GSC-ViT and DBCTNet achieve about 98% OA while maintaining relatively low Params of 104.976k and 40.265k, respectively. The proposed MSC-3DLT method, however, takes the lead in OA and Kappa, while also performing well in terms of Params and FLOPs.

For the XZ dataset, which contains 436 spectral channels, the challenge of feature extraction is exacerbated by spectral redundancies, increasing both Params and FLOPs. While HybridSN performs well on datasets with fewer spectral channels, it suffers from overfitting on the XZ dataset due to the large number of parameters. Transformer-based methods, however, excel in capturing sequential spectral information and, thus, achieve higher classification accuracy.

The HS2018 dataset presents great complexity due to its large number of material categories and the extremely uneven distribution of sample numbers. As a result, many methods struggle to achieve high classification accuracy for individual material categories. HybridKAN and LSSCM, while maintaining low FLOPs, demonstrate poor performance in terms of classification accuracy. DBCTNet benefits from its smaller number of channels, allowing it to achieve relatively favorable parameter amounts. The proposed MSC-3DLT method, however, achieves the highest classification accuracy in OA, AA, and Kappa and strikes a better balance between Params and FLOPs on this challenging dataset.

To qualitatively evaluate the classification performance of the proposed MSC-3DLT method, we conducted experimental comparisons with 11 SOTA methods. The complete visual classification results are shown in Figs. 8–11 on the SV, LK, XZ, and HS2018 datasets, respectively. By comparing similar land cover categories, our method produces classification maps that are closest to the ground truth land cover in terms of edges, boundaries, and noise distribution. In relatively simple regions with numerous samples, such as “Grapes untrained” and “Vineyard untrained” in the SV dataset, the CNN-based methods of RSSAN, LSSCM, and SPRN exhibit varying degrees of salt-and-pepper noises, with LSSCM showing the most severe noise. HybridSN performs well on the SV, LK, and HS2018 datasets by integrating multilayer convolutions for feature extraction. However, it exhibits overfitting on datasets with more spectral channels, such as XZ, leading to classification errors in large regions. HybridKAN produces significant classification errors across multiple datasets with small sampling rates. The CAN achieves relatively accurate classification results by combining CNNs and capsule networks, but the visual results are overly smooth in some land cover categories, resulting in the loss of details and misclassified boundaries. The methods of GATH,

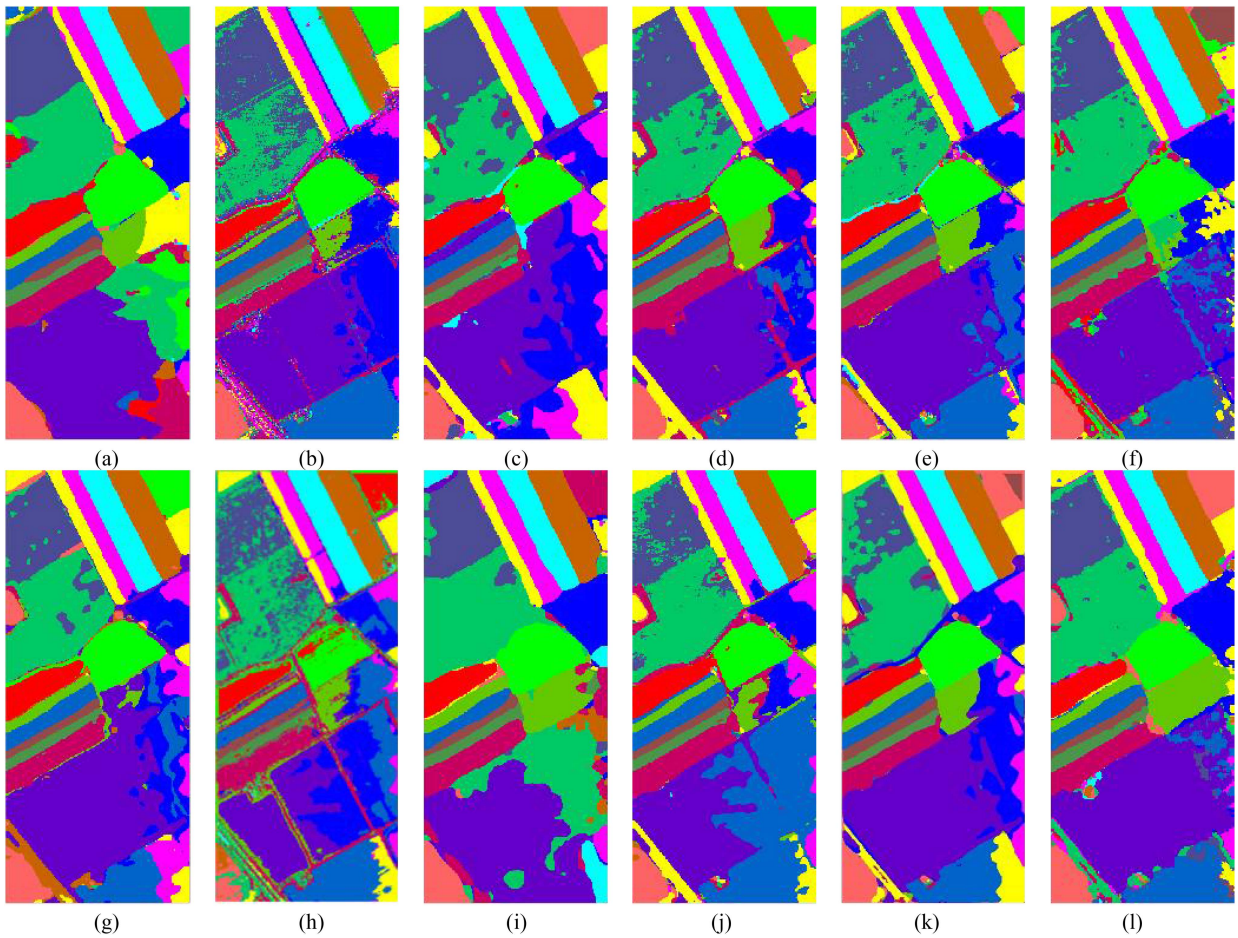


Fig. 8. Classification maps of different methods on the SV dataset. (a) HybridSN. (b) RSSAN. (c) LSSCM. (d) SPRN. (e) GAHT. (f) SSFTT. (g) MorphFormer. (h) HybridKAN. (i) CAN. (j) GSC-ViT. (k) DBCTNeT. (l) MSC-3DLT.

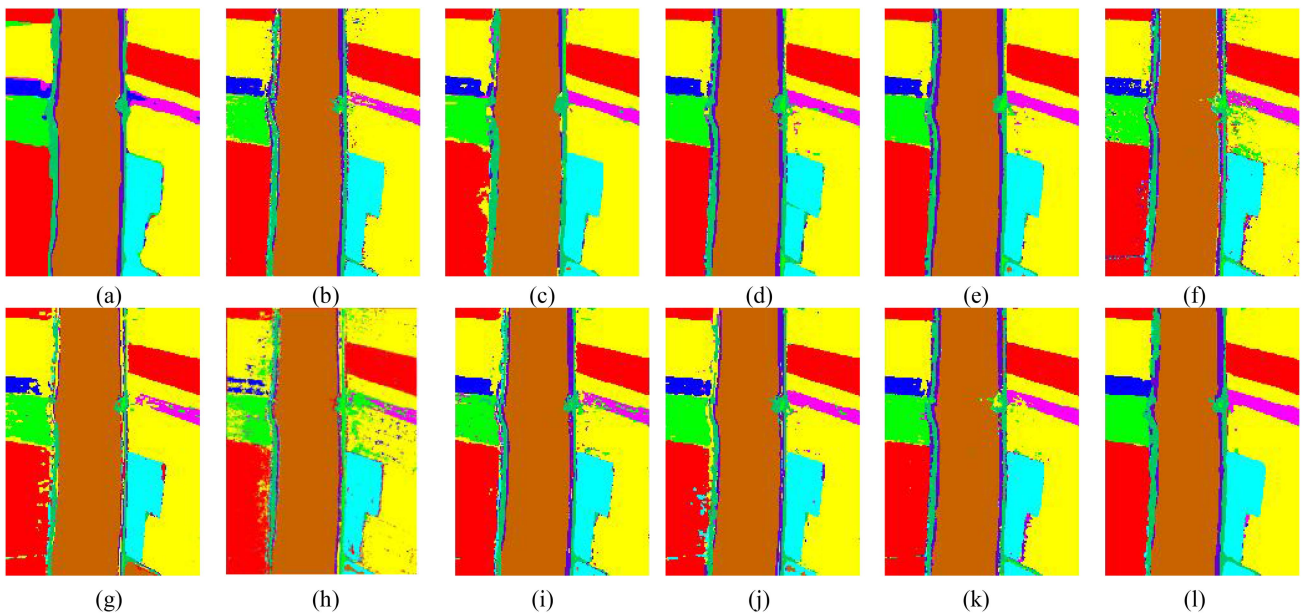


Fig. 9. Classification maps of different methods on the LK dataset. (a) HybridSN. (b) RSSAN. (c) LSSCM. (d) SPRN. (e) GAHT. (f) SSFTT. (g) MorphFormer. (h) HybridKAN. (i) CAN. (j) GSC-ViT. (k) DBCTNeT. (l) MSC-3DLT.

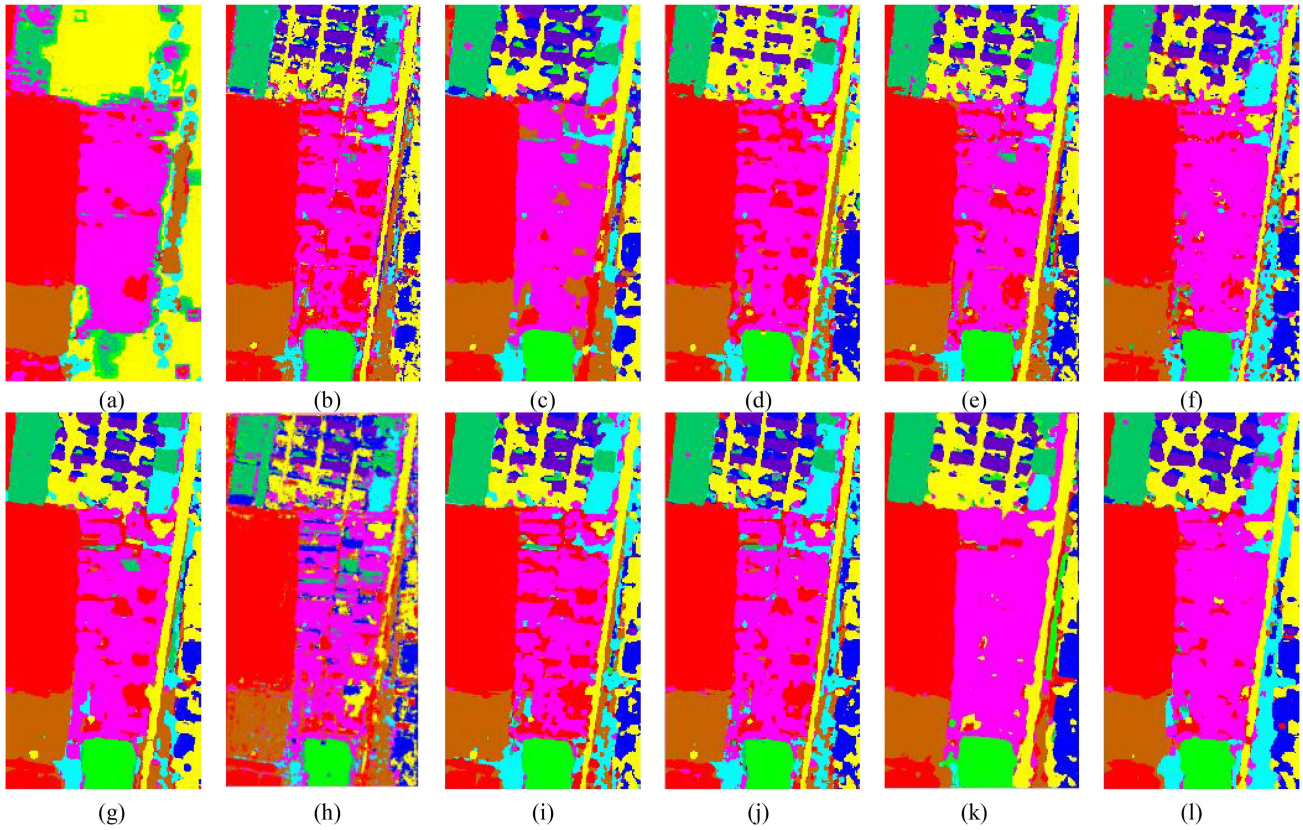


Fig. 10. Classification maps of different methods on the XZ dataset. (a) HybridSN. (b) RSSAN. (c) LSSCM. (d) SPRN. (e) GAHT. (f) SSFTT. (g) MorphFormer. (h) HybridKAN. (i) CAN. (j) GSC-ViT. (k) DBCTNeT. (l) MSC-3DLT.

SSFTT, MorphFormer, GSC-ViT, and DBCTNet achieve better visual classification results with minimum block noise by combining CNN and Transformer to simultaneously extract local spatial features and global spectral features.

Overall, the proposed MSC-3DLT method demonstrates competitive HSI classification performance across all four datasets. It achieves high classification accuracy at small sampling rates while requiring little computational resources, effectively handling complex scenarios and further verifying the robustness of our method.

E. Influence of Training Sample Size

To assess the generalization ability of the proposed MSC-3DLT method, we conduct a series of experiments across various sampling rates, with comparisons made to CNN- and transformer-based methods. For the SV and XZ datasets, the sampling rates of 0.1%, 0.3%, 0.5%, 0.7%, and 1% are adopted, while 0.05%, 0.1%, 0.2%, 0.5%, and 0.7% for LK dataset. The HSI classification results, as displayed in Fig. 12, reveal that our method performs well regardless of whether the sampling rate increases. Specifically, the OA of the MSC-3DLT consistently improves with higher sampling rates, indicating that the model demonstrates strong generalization ability. Notably, even at extreme sampling rates (e.g., there are as few as two training samples for categories such as "Lettuce romaine 4wk" and "Lettuce romaine 6wk" at a 1‰ sampling rate in the SV dataset), our MSC-3DLT method maintains high classification accuracy. This underscores the model's capacity to effectively leverage HSI data by jointly extracting spatial–spectral features, even with limited training samples. However, for the LK and XZ datasets, at sampling rates of 0.05% and 0.1%, the OA of the proposed MSC-3DLT is slightly lower than that of SPRN. This is largely due to SPRN's ability to capture detailed spatial features, utilizing its powerful local contextual modeling via convolution and attention, particularly in datasets with high spatial resolution.

Overall, the proposed MSC-3DLT method consistently delivers high classification performance in terms of OA across a wide range of sampling rates in all datasets.

F. Analysis on Learning Rate and Epochs

To verify the efficacy of the proposed MALR scheduler, we conduct experiments by comparing it with FLR and LDLR schedulers over multiple epochs. The initial learning rate for all three schedulers is set to 1e-2. The attenuation period for the LDLR scheduler is set to one-tenth of the maximum epochs with an attenuation factor of 0.5 [34], [37]. The overall classification accuracies, as depicted in Fig. 13, show that the proposed MALR consistently achieves the highest accuracy across all training epochs. This success is primarily attributed to the adaptive adjustment of the learning rate, which is based on the model fitness degree. In comparison, the FLR performs betters than the

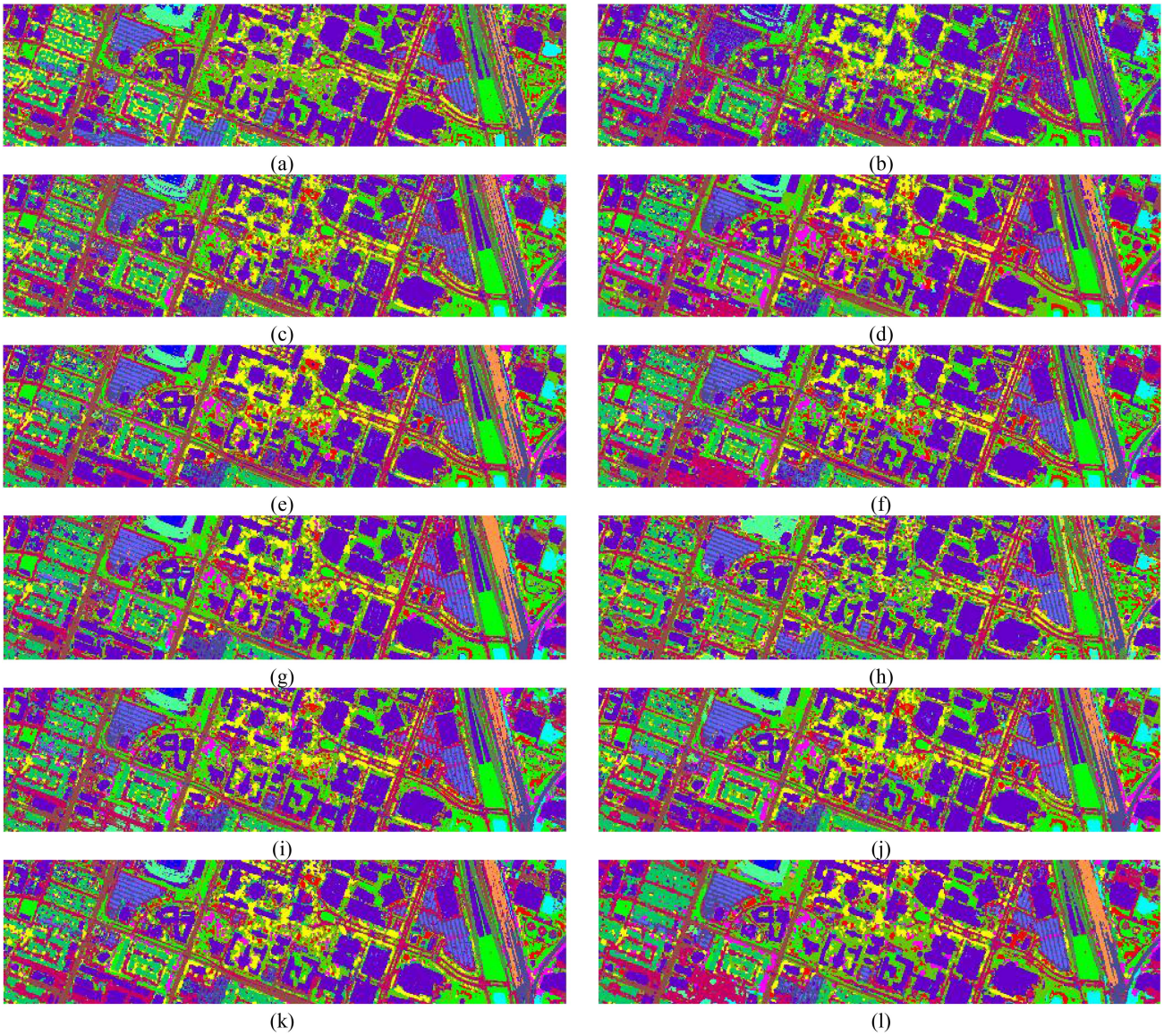


Fig. 11. Classification maps of different methods on the HS2018 dataset. (a) HybridSN. (b) RSSAN. (c) LSSCM. (d) SPRN. (e) GAHT. (f) SSFTT. (g) MorphFormer. (h) HybridKAN. (i) CAN. (j) GSC-ViT. (k) DBCTNeT. (l) MSC-3DLT.

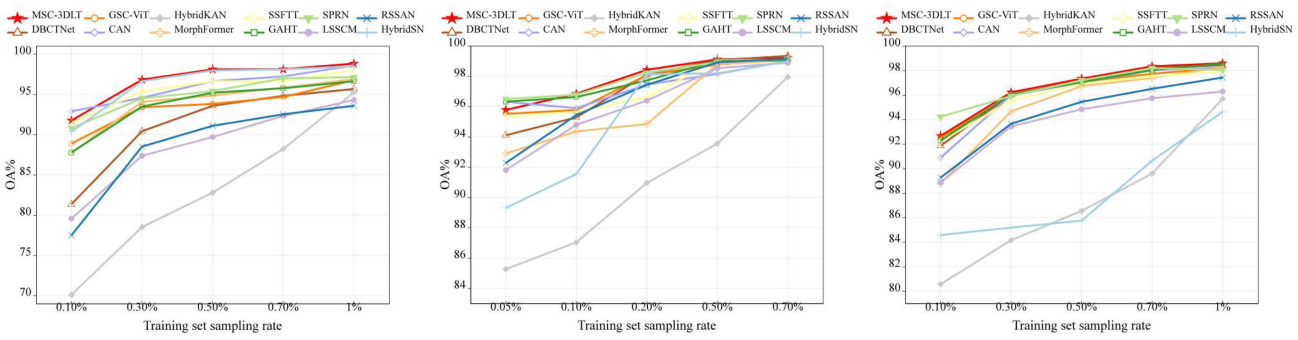


Fig. 12. Influence of different training sample rates in HybridSN, RSSAN, LSSCM, SPRN, GAHT, SSFTT, MorphFormer, HybridKAN, CAN, GSC-ViT, DBCTNet, and MSC-3DLT methods on classification accuracy. (a) SV dataset. (b) LK dataset. (c) XZ dataset.

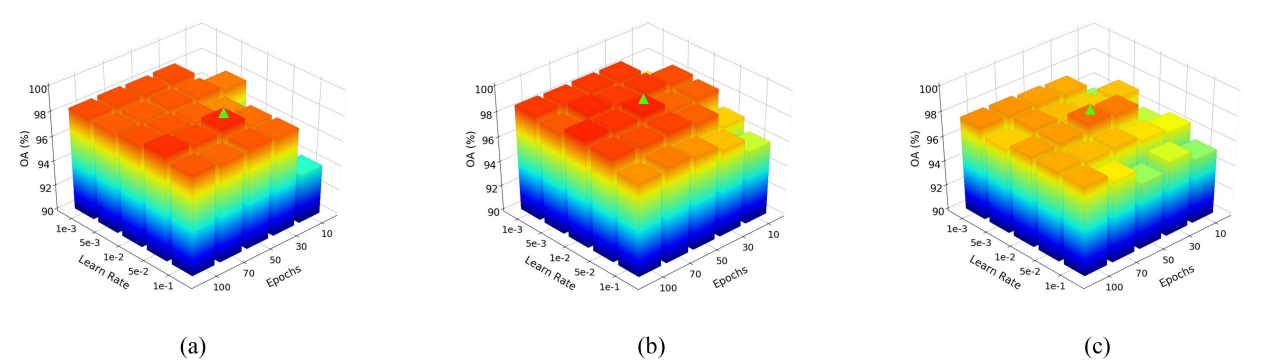


Fig. 13. Influence of different initial learning rates and training epochs on classification accuracy. (a) SV dataset. (b) LK dataset. (c) XZ dataset.

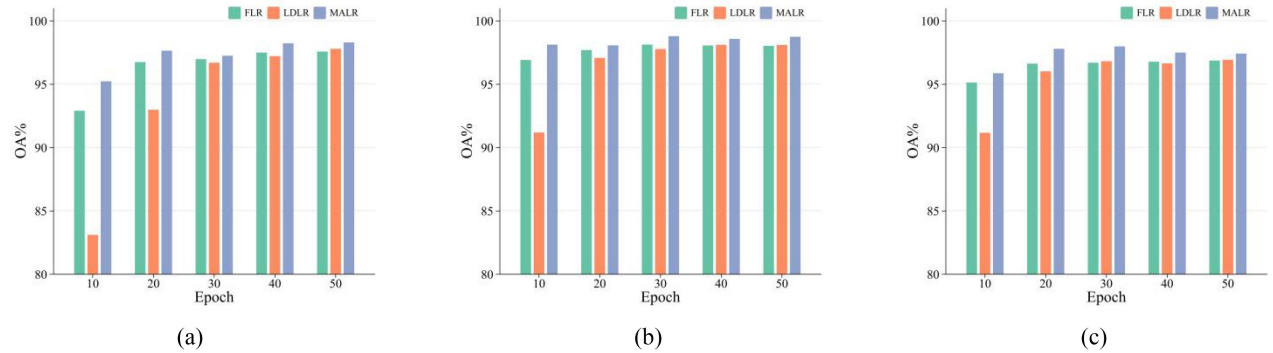


Fig. 14. Effects of different learning rate schedulers of FLR, LDLR, and MALR under different epochs to OA classification accuracy on three HSI datasets. (a) SV dataset. (b) LK dataset. (c) XZ dataset.

LDLR in the early epochs, as a larger learning rate is required for rapid model fitting during the initial training stages. The LDLR, on the other hand, decays the learning rate too quickly during the early phases. However, as the training progresses, the OA of the LDLR increases rapidly and eventually surpasses that of the FLR. This is because the increased number of epochs allows the LDLR and the FLR to sufficiently refine the model. Ultimately, during later stages of training, a more finely tuned learning rate is needed to optimize model fitting. The FLR, with its constant learning rate, tends to cause the model to oscillate around the optimal solution, whereas the MALR maintains an appropriate value throughout, accelerating model fitting in the early stages and adjusting the rate in later stages to guide the model toward its optimal solution.

In addition, we explored the impact of different initial learning rates and training epochs on the performance of the MALR. As shown in Fig. 14, the OA of our MSC-3DLT achieves high classification accuracy as early as 30 epochs and reaches the optimal solution by the 50th epoch. In contrast, most methods require 100 or more epochs to achieve similar results. This advantage is primarily due to the adaptive nature of the MALR strategy, which adjusts the learning rate based on real-time assessments of model fitting across multiple metrics. The ability of the MALR to make subtle adjustments to a few parameters allows the method to converge to the optimal solution quickly, thereby reducing the required training time and enhancing overall classification performance.

V. Discussion

The aforementioned experiments verify the effectiveness of the proposed MSC-3DLT in maintaining a lightweight characteristic and achieving higher classification accuracy at low sampling rates. It is primarily attributed to the MSC and 3DLT modules. The MSC module utilizes the multiscale convolution and channel shuffling to fully extract multigranular features and enhance feature interaction and fusion. The 3DLT module employs an improved LSPCA to replace traditional self-attention, extracting more detailed spatial–spectral features and reducing model complexity. Experiments demonstrate that the proposed MSC-3DLT achieves mean improvements in OA, AA, and Kappa of 2.20, 2.42, and 2.63, respectively, when compared to SOTA methods on the four HSI datasets of SV, LK, XZ, and HS2018.

However, although the proposed MSC-3DLT effectively preserves the inherent characteristics of HSI and improves data utilization, it still has several limitations. Some ground object features are difficult to effectively extract at low sampling rates due to sample imbalance or noise interference. For example, “Lettuce romaine 6wk” in the SV dataset, “Coals” in the XZ dataset, “Water” and “Unpaved parking lots” in the HS2018 dataset, etc., all have very few samples, making it challenging for the model to accurately learn and generalize these features. Moreover, “Mixed weed” in the LK dataset and “Cross walks” in the HS2018 dataset are highly susceptible to noise interference, resulting in feature confusion and difficulty in effective

distinction. The classification accuracy of these ground objects in the aforementioned two cases is significantly lower than other ground object categories. In the future, the model-based priors may be incorporated into models to better capture the intrinsic characteristics of HSI data and improve the classification performance of our methods.

VI. Conclusion

In this article, a novel MSC-3DLT method has been proposed for accurate and efficient HSI classification. Throughout the feature extraction process, the 3-D structure of HSI data is consistently maintained. In the early stages, the MSC module constructs feature pyramids and residual connections, enhancing the interaction between spatial and spectral features while maintaining spectral continuity by shuffling features belonging to different groups. In the later stages, the 3DLT module fuses spatial and spectral features from different dimensions, ensuring that the proposed MSC-3DLT method effectively deals with the inherent 3-D information, thereby greatly improving the utilization of the intrinsic characteristics of HSI cubic data. In addition, an adaptive learning rate has been proposed by monitoring multiple metrics during the model training process, substantially accelerating the training speed. The proposed method achieves better classification performance under limited training samples by comprehensively considering data, models, and training strategies.

References

[1] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

[2] A. Lobo, "Image segmentation and discriminant analysis for the identification of land cover units in ecology," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 5, pp. 1136–1145, Sep. 1997.

[3] F. Tsai and W. D. Philpot, "A derivative-aided hyperspectral image analysis system for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 2, pp. 416–425, Feb. 2002.

[4] M. Dalponte, H. O. Orka, T. Gobakken, D. Gianelle, and E. Naesset, "Tree species classification in boreal forests with hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2632–2645, May 2013.

[5] W. Huang, D. Zhou, L. Sun, Q. Chen, and J. Yin, "Adaptive pixel-level and superpixel-level feature fusion transformer for hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 16876–16889, 2024.

[6] A. Brook and E. B. Dor, "Quantitative detection of settled dust over green canopy using sparse unmixing of airborne hyperspectral data," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 9, no. 2, pp. 884–897, Feb. 2016.

[7] X. Yang and Y. Yu, "Estimating soil salinity under various moisture conditions: An experimental study," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2525–2533, May 2017.

[8] J.-P. Ardouin, J. Levesque, and T. A. Rea, "A demonstration of hyperspectral image exploitation for military applications," in *Proc. 10th Int. Conf. Inf. Fusion*, Jul. 2007, pp. 1–8.

[9] C. Ke, "Military object detection using multiple information extracted from hyperspectral imagery," in *Proc. Int. Conf. Prog. Informat. Comput.*, 2017, pp. 124–128.

[10] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[11] Q. Ye, P. Huang, Z. Zhang, Y. Zheng, L. Fu, and W. Yang, "Multiview learning with robust double-sided twin SVM," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 12745–12758, Dec. 2022.

[12] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[13] C. Cariou and K. Chehdi, "A new k-nearest neighbor density-based clustering method and its application to hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2016, pp. 6161–6164.

[14] S. Prasad and L. M. Bruce, "Limitations of principal components analysis for hyperspectral target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 625–629, Oct. 2008.

[15] G. Licciardi, P. R. Marpu, J. Chanussot, and J. A. Benediktsson, "Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 447–451, May 2012.

[16] Q. Ye, J. Yang, F. Liu, C. Zhao, N. Ye, and T. Yin, "L1-norm distance linear discriminant analysis based on an effective iterative algorithm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 114–129, Jan. 2018.

[17] L. Fu et al., "Learning robust discriminant subspace based on joint $L_{2,p}$ and $L_{2,s}$-norm distance metrics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 130–144, Jan. 2022.

[18] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[19] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sens.*, vol. 2015, pp. 1–12, Jan. 2015.

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

[21] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[22] H. Lee and H. Kwon, "Contextual deep CNN based hyperspectral classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 3322–3325.

[23] S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, "Learning sensor-specific spatial–spectral features of hyperspectral images via convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4520–4533, Aug. 2017.

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

[25] M. He, B. Li, and H. Chen, "Multi-scale 3D deep convolutional neural network for hyperspectral image classification," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 3904–3908.

[26] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

[27] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[28] L. Sun, C. Ma, H. J. Shim, Z. Wu, and B. Jeon, "Adjacent superpixel-based multiscale spatial–spectral kernel for hyperspectral classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1905–1919, Jun. 2019.

[29] H. Gao, Y. Yang, C. Li, L. Gao, and B. Zhang, "Multiscale residual network with mixed depthwise convolution for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3396–3408, Apr. 2021.

[30] R. Shang, H. Chang, W. Zhang, J. Feng, Y. Li, and L. Jiao, "Hyperspectral image classification based on multiscale cross-branch response and second-order channel attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532016.

[31] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.

[32] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

[33] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021.

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

[35] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715.
[36] J. Zhang, Z. Meng, F. Zhao, H. Liu, and Z. Chang, "Convolution transformer mixer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6014205.
[37] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539014.
[38] Q. Yu, W. Wei, D. Li, Z. Pan, C. Li, and D. Hong, "HyperSINet: A synergetic interaction network combined with convolution and transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5508118.
[39] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.
[40] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral–spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503615.
[41] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
[42] Z. Li, Z. Xue, Q. Xu, L. Zhang, T. Zhu, and M. Zhang, "SPFormer: Self-pooling transformer for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5502019.
[43] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[44] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–12.
[45] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 3965–3977.
[46] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, "Biformer:Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10323–10333.
[47] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "Efficientvit:Memory efficient vision transformer with cascaded group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14420–14430.
[48] Y. Fang, Q. Ye, L. Sun, Y. Zheng, and Z. Wu, "Multiattention joint convolution feature representation with lightweight transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513814.
[49] Z. Zhao, X. Xu, S. Li, and A. Plaza, "Hyperspectral image classification using groupwise separable convolutional vision transformer network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5511817.
[50] X. Zhang, Y. Su, L. Gao, L. Bruzzone, X. Gu, and Q. Tian, "A lightweight transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5517617.
[51] M. Guo, C. Lu, Q. Hou, Z. Liu, M. Cheng, and S. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 1140–1156.
[52] Z. Xiong, Y. Yuan, and Q. Wang, "AI-NET: Attention inception neural networks for hyperspectral image classification," in *Proc. lEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2647–2650.
[53] M. Han, R. Cong, X. Li, H. Fu, and J. Lei, "Joint spatial-spectral hyperspectral image classification based on convolutional neural network," *Pattern Recognit. Lett.*, vol. 130, pp. 38–45, 2020.
[54] J. Feng, L. Wang, H. Yu, L. Jiao, and X. Zhang, "Divide-and-conquer dual-architecture convolutional neural network for classification of hyperspectral images," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 484.
[55] K. Safari, S. Prasad, and D. Labate, "A multiscale deep learning approach for high-resolution hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 18, no. 1, pp. 167–171, Jan. 2021.
[56] H. Gong et al., "Multiscale information fusion for hyperspectral image classification based on hybrid 2d-3d CNN," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2268.
[57] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 22–31.
[58] C. Wu, L. Tong, J. Zhou, and C. Xiao, "Spectral–spatial large kernel attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5508814.
[59] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 10809–10819.
[60] W. Yu et al., "MetaFormer baselines for vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 896–912, Feb. 2024.
[61] Z. Meng, L. Jiao, M. Liang, and F. Zhao, "A lightweight spectral-spatial convolution module for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5505105.
[62] X. Zhang, S. Shang, X. Tang, J. Feng, and L. Jiao, "Spectral partitioning residual network with spatial attention mechanism for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5507714.
[63] A. Jamali et al., "How to learn more? Exploring Kolmogorov–Arnold networks for hyperspectral image classification," *Remote Sens.*, vol. 16, no. 21, 2024, Art. no. 4015.
[64] N. Wang et al., "Capsule attention network for hyperspectral image classification," *Remote Sens.*, vol. 16, no. 21, 2024, Art. no. 4001.
[65] R. Xu, X.-M. Dong, W. Li, J. Peng, W. Sun, and Y. Xu, "DBCTNet: Double branch convolution-transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5509915.

Qinggang Wu received the Ph.D. degree in computer science from Dalian Maritime University, Dalian, China, in 2012.

He is currently an Associate Professor with the College of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou, China. His current research interests include remote sensing image processing, hyperspectral image classification, artificial intelligence, and deep learning.

Mengkun He received the B.S. degree in computer science and technology from the Henan University of Urban Construction, Pingdingshan, China, in 2022. He is currently working toward the M.E. degree in computer technology with the Zhengzhou University of Light Industry, Zhengzhou, China.

His research interests include hyperspectral image classification, machine learning, and deep learning.

Qiqiang Chen was born in Henan, China, in 1984. He received the Ph.D. degree in radio physics from Lanzhou University, Lanzhou, China, in 2015.

He is currently a Lecturer with the Zhengzhou University of Light Industry, Zhengzhou, China. His current research interests include hyperspectral classification, image processing, and machine learning.

Le Sun (Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 2009 and 2014, respectively.

From 2015 to 2018, he was a Postdoctoral Researcher with the School of Electronic and Electrical Engineering, Sungkyunkwan University Natural Science Campus, Suwon, South Korea. He is currently a full Professor with the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing. His research interests include high-dimensional signal processing and deep learning.

Chao Ma was born in Henan, China, in 2000. He is working toward the M.S. degree in big data technology and engineering with the Zhengzhou University of Light Industry, Zhengzhou, China.

His research interests include hyperspectral image processing.