MSLKCNN: a simple and powerful multi-scale large kernel CNN for hyperspectral image classification.

LIU, X., NG, A.H.-M., LEI, F., REN, J., GUO, L. and DU, Z.

2025

MSLKCNN: A Simple and Powerful Multi-scale Large Kernel CNN for Hyperspectral Image Classification

Xun Liu, Alex Hay-Man Ng, *Senior Member, IEEE*, Fangyuan Lei, *Member, IEEE*, Jinchang Ren, *Senior Member, IEEE*, Li Guo, Zheyuan Du

***Abstract*—Deep learning-based hyperspectral image (HSI) classification models typically utilize multiple feature extraction layers to learn the features of land covers. Nevertheless, they encounter challenges, e.g., 1) Transformers require substantial computational resources, and 2) these layers are carefully assembled and designed. Recently, large kernel convolutional neural networks (LKCNNs) show excellent performance in natural visual tasks. To tackle these limitations and explore the capability of LKCNNs for HSI classification, we present a novel simple and powerful multi-scale large kernel convolutional neural network architecture (MSLKCNN) with the largest kernel size as large as 15×15, in contrast to commonly used 3×3, for HSI classification. MSLKCNN avoids these specialized designs, comprising a noise suppression module (NSM) and a multi-scale large kernel convolution (MSLKC). Specifically, NSM is first used to suppress the noise and reduce the number of the bands before extracting the features. Then, MSLKC, as the only feature extraction layer of MSLKCNN, joints three parallel convolutions to capture the features of various types (i.e. spectral, spectral-spatial) and ranges (i.e., small local, larger local, and global) from the dimension of scale: (C1) convolution with a kernel size of 1×1 is used to extract spectral features; (C2) multi-scale large kernel depthwise separable convolution (MLKDC) is proposed to learn the spectral-spatial features of different ranges including short-range, middle-range, and long-range; and (C3) multi-scale dilated depthwise separable convolution (MDDC) is designed to aggregate the spectral-spatial features between land covers at various distances. Extensive experimental results on three public HSI datasets demonstrate the competitiveness of the proposed MSLKCNN compared with several state-of-the-art methods.**

***Index Terms*—Hyperspectral image (HSI) classification, convolutional neural network (CNN), multi-scale convolution, large kernel convolution.**

Manuscript received 2024; revised 2024.

This work was supported by the National Natural Science Foundation of China (grant number 42274016), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (grant number 2019ZT08L213), and the Guangdong Provincial Key Laboratory of Intellectual Property & Big Data (No. 2018B030322016). *(Corresponding author: Alex Hay-Man Ng; Li Guo.)*

Xun Liu is with the School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China (e-mail: liuxun.stf@gmail.com).

Alex Hay-Man Ng is with the School of Civil and Transportation Engineering, Guangdong University of Technology, Guangzhou 510006, China (e-mail: hayman.ng@gdut.edu.cn).

Fangyuan Lei is with the Guangdong Provincial Key Laboratory of Intellectual Property Big Data, Guangdong Polytechnic Normal University, Guangzhou 510665, China (e-mail: leify@gpnu.edu.cn).

Jinchang Ren is with the National Subsea Centre, Robert Gordon University, Aberdeen AB21 0BH, U.K. (e-mail: jinchang.ren@ieee.org).

Li Guo is with the Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing 100048, China (e-mail: guol@lasac.cn).

Zheyuan Du is with the Geoscience Australia, Canberra 2609, Australia, and also with the School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: z.du@unsw.edu.au).

I. Introduction

HYPERSPECTRAL image (HSI) consists of hundreds of spectral bands obtained by hyperspectral remote sensors. It contains abundant spectral-spatial information, and is more effective and accurate in classifying land cover types than RGB images. Because of these advantages, HSI has been widely applied to various applications, ranging from environmental monitoring to geological exploration, medical diagnosis, and object tracking [1]–[5]. HSI classification, which classifies each pixel into a certain label, plays a deterministic role in the applications [6]–[9].

Over the last few decades, a variety of methods have been explored and proposed for HSI classification, which can be summarized as traditional machine learning methods and deep learning methods. The traditional methods such as decision tree [10], K-nearest neighbor classifier [11], linear regression (LR) [12], random forest [13], support vector machine (SVM) [14], sparse representation [15], wavelets [16], and morphological profiles [17] are utilized to extract the spectral–spatial features contained in HSI. However, these methods are based on the handcrafted features that heavily depend on professional expertise and are empirical [18], so they are difficult to learn the robust deep feature representations of HSI.

Recently, inspired by the promising results of deep learning (DL) in various fields, DL models have been applied to HSI classification. Unlike the traditional methods, DL models can learn robust deep feature representations automatically without handcrafted feature engineering [8]. DL methods, such as stacked autoencoders (SAEs) [19], recurrent neural networks (RNNs) [6], [20], convolutional neural networks (CNNs) [21]–[25], capsule networks (CapsNets) [26], graph convolutional networks (GCNs) [27]–[29], and Transformer [7], [30] have been explored for HSI classification. Among them, CNN-, GCN-, and Transformer-based models have received more attention.

CNN-based Models: 1D-CNN-based model [31] is designed to learn the spectral information. SSFC [32], utilizing 2D-CNN, is employed to extract the spatial-spectral features. Compared to several DL models, Li et al. [23] better exploit the discrimination information via 3D-CNN with fewer parameters. Besides, channel-based CNN methods including single-channel CNN [32], dual-channel CNN [24], and multi-channel CNN [33], have been designed to learn the hierarchical high-level features. To overcome the degradation [34] in deeper CNNs, SSRN [35] and FDSSC [36] introduce residual connection [34] and dense connection [37] to CNNs for classifying hyperspectral data, respectively. However, most of these models only extract single-scale features from the fixed-size image patches. In complex HSI with limited training samples, the classification results of these single-scale models may be suboptimal and unstable [38], [39]. To address these limitations, several multi-scale CNN models have been proposed to extract multi-scale features by using multi-scale filter banks. Lee et al. [40] present a multi-scale filter bank in the first layer of their network, which is constructed using three parallel convolutions with different kernel sizes to extract multi-scale features. Gong et al. [41] explore three multi-scale CNN (MS-CNNs) to enhance the learning ability of single-scale CNNs for HSI classification. To consider the complementary and related information among features of various scales, Li et al. [42] develop a multi-scale deep CNN with residual blocks to extract multi-scale spatial-spectral features from input image patches of diverse scales. To learn more pixel-level discriminative features, attention-based multi-scale CNN methods such as DBMA [43] and DBDA [44] have been introduced. Ma et al. [43] propose a double-branch multi-attention network (DBMA) that captures more discriminative spatial-spectral features through the designs of two branches and an attention mechanism. To enhance the performance of several networks such as DBMA, Li et al. [44] present a double-branch dual-attention network (DBDA) by developing an attention mechanism. In DBDA, two branches are constructed to optimize spatial-spectral features: a channel attention block and a spatial attention block. Additionally, Wang et al. [45] introduce an attention-based multi-scale CNN framework utilizing two designs: a multi-scale spatial-channel attention mechanism and a shuffle block. Although these CNN-based models have advantages in extracting local spatial-spectral features, they struggle to directly capture the relationships of land covers at medium- and long-term distances [30], [54], [65], due to their limited receptive fields.

GCN-based Models: Compared to CNNs that tend to capture local features, graph convolutional networks (GCNs) [46] and their variant models are capable of modeling remote dependencies among various nodes in space, overcoming the problems of these CNN-based models. Therefore, GCNs have become a hot topic and have been applied to learn the features of HSI. Qin et al. [27] propose spectral–spatial GCNs (S2GCNs) to extract spatial-spectral features, modeling each pixel of HSI as a graph node. However, S2GCNs suffer from the concerns of high complexity. To assign appropriate weights to neighboring nodes, Sha et al. [28] design graph attention networks (GATs) via the spatial-spectral similarity of pixels. By utilizing the technologies such as dense connections [37] and dilated convolutions [47], Bai et al. [29] develop an attention framework termed DAGCN, which constructs deep graph convolutional networks (DeepGCNs) to consider the non-Euclidean features of HSI. These graph networks regard each pixel in HSI as a graph node, which requires significant computing and time resources. To avoid these limitations, many superpixel-based GCNs [9], [48]–[53] have been explored. For instance, Wan et al. [48] and Ding et al. [49] reduce the number of nodes by constructing superpixels instead of using individual pixels as graph nodes. Nevertheless, in these superpixel-based models, the features of each superpixel are assumed to be uniform, ignoring the individual characteristics of pixels. Consequently, the classification performance is limited. To tackle these issues, several fusion networks that integrate CNNs and GCNs have been proposed, aiming to supplement the shortcomings of their respective networks while leveraging their advantages [54]–[57]. However, these fusion models may necessitate several elaborate designs, including the number of layers, the number of filters for each layer, and the fusion scheme of different networks.

Transformer-based Models: Recently, inspired by the remarkable performance of vision Transformers (ViTs) [58] in natural image processing, Transformer-based models [7], [30], [59]–[61] have been investigated for HSI classification, achieving excellent classification results due to their powerful ability to model global contextual information. He et al. [59] introduce a Transformer-based method called HSI-BERT, which utilizes the multihead self-attention (MHSA) [62] mechanism to capture the relationships between land covers at long-term distances. Hong et al. [30] develop a Transformer-based model, SpectralFormer, that extracts locally spectral features from multiple neighboring bands and conveys memory-like components from shallow to deep layers. Subsequently, SSFTT [7] employs a Gaussian weighted feature tokenizer and transformer encoder module to extract spatial-spectral features. While these models effectively establish long-term dependencies in HSI by utilizing Transformers instead of traditional convolutions with shape-fixed kernel, they may still face two limitations: 1) they may overlook the contribution of local neighboring bands to object classification [63] owing to the abundant spectral bands contained in HSI; and 2) these models may struggle to extract local features effectively. To address these limitations, several fusion architectures have been explored. For instance, MVAHN [64] and GTFN [65] combine the strengths of GCN and Transformer, utilizing GCN to learn the contextual information of classified pixels, while leveraging Transformer to model long-term dependencies among these pixels. Subsequently, Zhao et al. [66] propose a hybrid approach that integrates CNN with Transformer to capture both local and global features. Nevertheless, these Transformer-based models suffer from a quadratic computational complexity of Transformer, making them impractical when dealing with large HSI with numerous labelled pixels. Furthermore, similar to these GCN-based fusion networks, these Transformer-based models also require customized designs.

In summary, most CNN, GCN, and Transformer models typically increase the size of their receptive fields by stacking layers from their respective networks. To further enlarge the size of the receptive field and improve learning ability, a range of techniques such as residual connections and attention mechanisms are used to facilitate deeper structures and better adjust the contribution of pixel features. However, each network type typically encounters its own specific limitations: 1) CNN-based networks have difficulties in extracting the relationships between land covers at medium- and long-term distances, owing to the local connectivity of these CNN models; 2) many GCN-

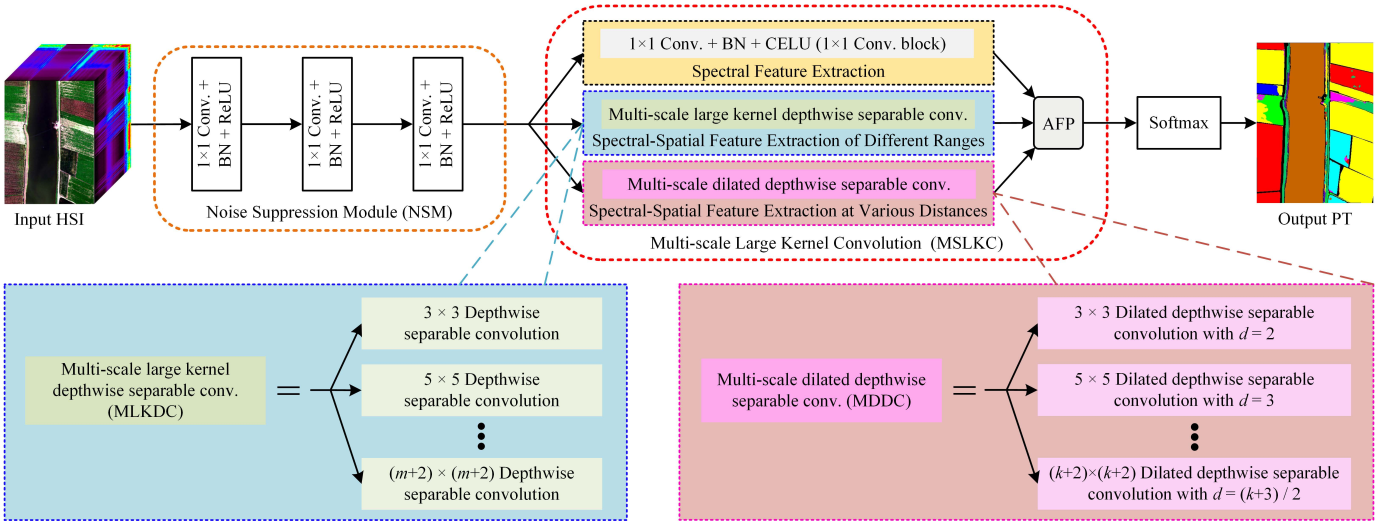
Input HSI
1×1 Conv. + BN + ReLU
1×1 Conv. + BN + ReLU
1×1 Conv. + BN + ReLU
Noise Suppression Module (NSM)
1×1 Conv. + BN + CELU (1×1 Conv. block)
Spectral Feature Extraction
Multi-scale large kernel depthwise separable conv.
Spectral-Spatial Feature Extraction of Different Ranges
Multi-scale dilated depthwise separable conv.
Spectral-Spatial Feature Extraction at Various Distances
Multi-scale Large Kernel Convolution (MSLKC)
AFP
Softmax
Output PT
Multi-scale large kernel depthwise separable conv. (MLKDC)
=
3 × 3 Depthwise separable convolution
5 × 5 Depthwise separable convolution
(m+2) × (m+2) Depthwise separable convolution
Multi-scale dilated depthwise separable conv. (MDDC)
=
3 × 3 Dilated depthwise separable convolution with d = 2
5 × 5 Dilated depthwise separable convolution with d = 3
(k+2)×(k+2) Dilated depthwise separable convolution with d = (k+3) / 2

Fig. 1: Architecture of the proposed MSLKCNN, which consists of two mainly components: a noise suppression module (NSM) to reduce the noise and bands, and a multi-scale large kernel convolution (MSLKC) to capture the features of different types (i.e. spectral, spectral-spatial) and ranges (i.e. small local, larger local, and global). MSLKC includes a 1×1 convolution block, MLKDC, MDDC, and average fusion pooling (AFP).

based networks tend to overlook the individual features of pixels by operating on superpixel-based nodes; and 3) Transformer-based networks often suffer from computational inefficiency. Moreover, these models necessitate a series of customized and intricate designs such as the number of layers, the filter numbers and sizes for each layer, the way of residual connections, and the implementation of attention mechanisms. These designs heavily rely on the expertise, experience, and time of the researcher.

Recently, large kernel CNNs (LKCNNs) [67]–[71] have come into focus. These LKCNNs leverage their large kernel convolutions to expand the receptive field, thereby demonstrating a superior capability in extracting long-term (global) features compared to traditional CNNs. Although these LKCNNs have achieved promising performance in natural visual tasks, their potential for high-dimensional visual domains such as HSI classification has not been fully explored.

To address these limitations that exist in the CNN-, GCN-, and Transformer-based models and explore the potential of LKCNNs in HSI classification, we introduce a novel end-to-end multi-scale large kernel CNN (MSLKCNN) for HSI classification (Fig. 1). MSLKCNN circumvents these elaborate designs described in the aforementioned models while scaling up the largest kernel size to 15×15. Specifically, we first develop a noise suppression module (NSM) to suppress the noise and reduce the bands in raw HSI. Subsequently, a new multi-scale large kernel convolution (MSLKC), consisting of three parallel convolution components, is introduced to extract comprehensive features of various types and scales. Among the three components, the 1×1 convolution is employed to extract spectral features, the multi-scale large kernel depthwise separable convolution (MLKDC) is utilized to learn short-range (small local), medium-range (larger local), and long-range (global) spectral-spatial features, and the multi-scale dilated depthwise separable convolution (MDDC) focuses on capturing spectral-spatial features between land covers at diverse distances. Finally, we introduce an average fusion pooling (AFP) to fuse the comprehensive features extracted by the three components. The main contributions of this article are summarized as follows:

1) We design a novel MLKDC composed of a series of arithmetic depthwise separable convolutions (DSCs) [72] with a similar topology in a parallel manner to capture the features in regions of different sizes.

2) We develop a new type of convolution, MDDC, by combining DSC with dilated convolution [47], to establish the relationships between land covers at various distances.

3) Based on our MLKDC and MDDC, we propose a novel MSLKC capable of extracting discriminative features across diverse types and ranges. Leveraging the proposed MSLKC, we introduce MSLKCNN, a simple and powerful architecture designed to effectively learn spectral-spatial features and circumvent the need for the carefully and customized designs of existing networks.

The rest of the paper is organized as follows. Section II introduces the proposed MSLKCNN. We evaluate the performance of MSLKCNN in Section III. In Section IV, we summarize the paper.

II. PROPOSED METHOD

An overview of the proposed MSLKCNN is illustrated in Fig. 1, which includes three modules: 1) NSM that eliminates noise and reduce the number of bands in raw HSI; 2) MSLKC that leverages three parallel convolutions to extract comprehensive features of different types and scales; and 3) a Softmax classification module that predicts the label for each pixel. In the following sections, we will describe the details of each module.

A. Noise Suppression Module (NSM)

Original HSI contains redundant band information and is subject to noise. To address these limitations, we present a new noise suppression module (NSM). In our NSM, we deal with the HSI using three successive 1×1 convolutions with a few filters, batch normalization (BN) techniques, and ReLU functions to suppress noise and reduce the number of bands.

Let X^l be the input feature map of the l-th convolutional layer, for the spatial location $p_1 = (x, y)$, the output feature map of the l-th 1×1 convolutional layer in the i-th spectral channel, denoted as $X_i^{l+1}(p_1)$, can be written as

$$X_i^{l+1}(p_1) = \text{ReLU}(\text{BN}(W_i^l \cdot X_i^l(p_1) + b_i^l)) \,, \tag{1}$$

where $X_i^l(p_1)$ denotes the value of the l-th 1×1 convolutional layer in the i-th input spectral channel at the location p_1, W_i^l denotes the trainable weight of the i-th kernel with a size of 1×1 in the l-th convolutional layer, and b_i^l is the bias of the i-th kernel in the l-th convolutional layer.

B. Multi-scale Large Kernel Convolution (MSLKC)

In this section, we propose a novel multi-scale large kernel convolution (MSLKC) that scales up the largest kernel size to 15×15, as depicted in Fig. 1. MSLKC comprises three convolutions and one fusion pooling operation: 1) a 1×1 convolution block, which is used to capture spectral features; 2) a multi-scale large kernel depthwise separable convolution (MLKDC) that extracts local and global spectral-spatial features of HSI in parallel; 3) a multi-scale dilated depthwise separable convolution (MDDC) that can learn the spatial relationships between pixels at different distances without additional parameters and the reduction of image resolution; and 4) an average fusion pooling (AFP) that fuses the features extracted by the three convolutions. In the subsequent sections, we will describe the four components in detail.

1) 1×1 Convolution Block: Our 1×1 convolution block, as one of the components of MSLKC, consisting of a 1×1 convolution, a BN, and an activation function CELU [73], is applied to learn spectral features. Let X_N be the features transferred by NSM, then the 1×1 convolution block can be expressed as

$$H_1 = \text{ReLU}(\text{BN}(\text{Conv1}(X_N))) \,, \tag{2}$$

where Conv1 denotes the 1×1 convolution, and H_1 is the output features of the 1×1 convolution block.

2) Multi-scale Large Kernel Depthwise Separable Convolution (MLKDC): Since depthwise separable convolution (DSC) [72] significantly reduces the number of computations and parameters compared to ordinary convolution, DSC is preferred over ordinary convolution for extracting features from HSI [74]. Fig. 2 illustrates the diagrams of an ordinary 3×3 convolution, DSC, and the proposed MLKDC. DSC decomposes the 3×3 convolution into a 3×3 depthwise convolution (DWC) [72] and a pointwise convolution (PWC) [72], aiming to significantly reduce the number of parameters and calculations. The DWC convolves each channel of input feature map separately, which is equivalent to a 3×3 group convolution [75] with the number of groups equal to the number of input channels. The PWC is a 1×1 convolution that learns the channel correlations among all feature maps extracted by the DWC operation. The advantages of this DSC motivate us to introduce it into our model for efficient computation.

To achieve the multi-scale feature extraction of HSI, we propose a novel MLKDC by designing arithmetic DSCs (DSCs with various kernel sizes in an arithmetic progression), as illustrated in Fig. 2 (c). The proposed MLKDC consists of $\frac{m+1}{2}$ parallel DSCs. These convolutions share the same input features learned by NSM, perform convolution operations in an equal-width manner, maintain a similar topology, and are constrained by two simple rules: (i) apart from the different spatial sizes of these DWCs, all other hyperparameters (filter numbers, strides) remain the same; and (ii) the spatial sizes are distributed in an arithmetic progression with common difference of 2. The equal-width manner and the first rule ensure that the sizes of the output and input feature maps are the same for each parallel DSC. Based on these two rules, we only need to design the first template DSC and set the scale numbers, and the proposed MLKDC can be determined accordingly. Consequently, these two rules greatly simplify the design space, enabling us to focus on a limited number of hyperparameters.

In every DSC, we utilize PWC (i.e., 1×1 convolution) for spectral feature extraction, followed by the extraction of spatial features through DWC. For the designed DWCs, the spatial sizes are evenly distributed, ranging from 3×3 to $(m+2) \times (m+2)$, where $m+2$ represents the largest kernel size within MLKDC. Since we set $m+2$ to a large value of 15 (i.e., $m = 13$) in our experiments, the proposed MLKDC is a large kernel convolution. We take X_N as input, then its output feature map H_{MLKDC} can be computed as:

$$H_{\text{MLKDC}} = \text{MLKDC}(X_N) = \{\tilde{H}_3, \tilde{H}_5, \ldots, \tilde{H}_{m+2}\} \,, \tag{3}$$

where $\tilde{H}_{m+2}$ denotes the extracted features from X_N by $(m+2) \times (m+2)$ DSC.

In the proposed MLKDC, we use small kernel convolutions (3×3 and 5×5) to extract local features, convolutions with medium-sized kernels (e.g., 7×7) to capture larger local features, and large kernel convolutions (e.g., 15×15) to learn global features.

3) Multi-scale Dilated Depthwise Separable Convolution (MDDC): In images, dilated convolution [47] expands the receptive field without adding extra parameters by sampling pixels in the neighborhood at intervals compared with ordinary convolution, and has been successfully applied to HSI [76]. This motivates us to leverage the benefits of dilated convolution in designing our model. To extract features at different scales in HSI, we design a new MDDC that consists of $\frac{k+1}{2}$ parallel dilated depthwise separable convolution (DDSC) modules (Fig. 3). All modules carry out convolution operations in an equal-width convolutional way, maintain a similar structure, and are subject to two simple rules: (i) except for the various spatial sizes and dilation factors of these DDSCs, the hyperparameters (filter numbers, strides) are shared among these DDSCs; and (ii) spatial sizes and dilation factors are arranged in arithmetic progressions with common differences of 2 and 1, respectively. The equal-width way and the first rule ensure that the sizes of the output and input feature maps are the same for each

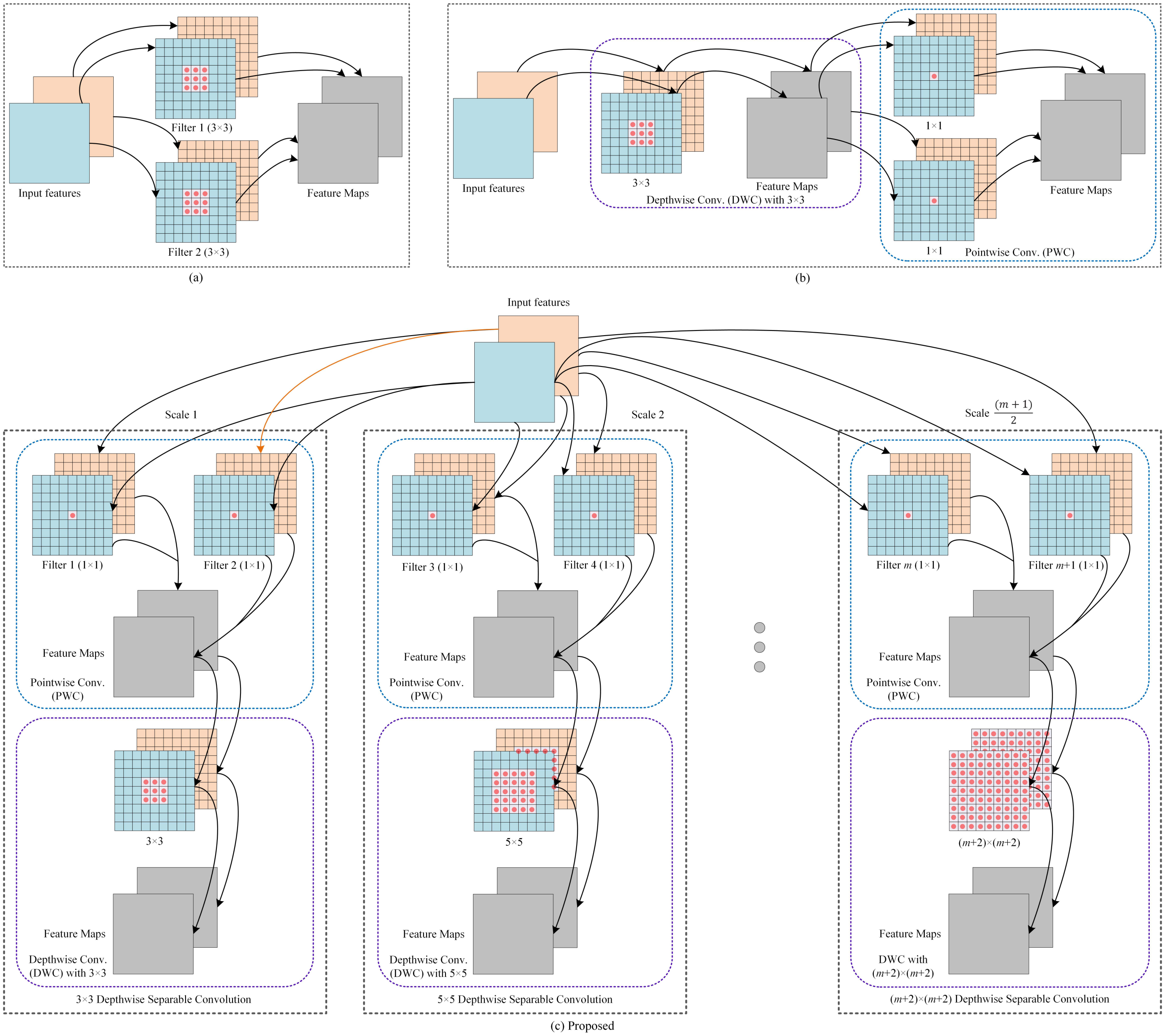

Fig. 2: Various convolutions. (a) Ordinary convolution. (b) Depthwise separable convolution. (c) Proposed multi-scale large kernel depthwise separable convolution (MLKDC).

parallel module. Similar to the proposed MLKDC, with the two rules, we only need to design the first template module and set the scale numbers, and the proposed MDDC can be determined accordingly. Therefore, the design of MDDC is quite straightforward.

As depicted in Fig. 3, DDSC represents a modified dilated convolution that combines DSC with dilated convolution, and can also be viewed as a new DSC with a dilation factor d. In every DDSC, PWC and DDC are utilized to learn the spectral and spatial features, respectively. For the proposed MDDC, the spatial sizes are evenly distributed, ranging from 3×3 to $(k+2) \times (k+2)$, where $k+2$ denotes the largest kernel size in MDDC. Let H_{MDDC} be the output features of the proposed MDDC, then we have

$$H_{\text{MDDC}} = \text{MDDC}(X_N) = \{\hat{H}_3, \hat{H}_5, \dots, \hat{H}_{k+2}\}, \quad (4)$$

where $\hat{H}_{k+2}$ is the learned features from X_N by $(k+2) \times (k+2)$ DDSC.

In our MDDC, we use parallel DDSCs with various kernel sizes and dilation factors to learn the multi-scale features of HSI in parallel and build relationships between pixels at different distances.

4) Average Fusion Pooling (AFP): In the proposed MSLKC, we design three convolutions to extract the features of HSI in parallel: a 1×1 convolution block, MLKDC, and MDDC. By fusing these features extracted by the three convolutions, the strengths of the three various convolutions can be effectively combined. In HSI classification task, feature fusion schemes commonly include column concatenation fusion [55] and sum fusion [64]. We explore a new fusion scheme for this task, termed average fusion pooling (AFP), which aims to integrate

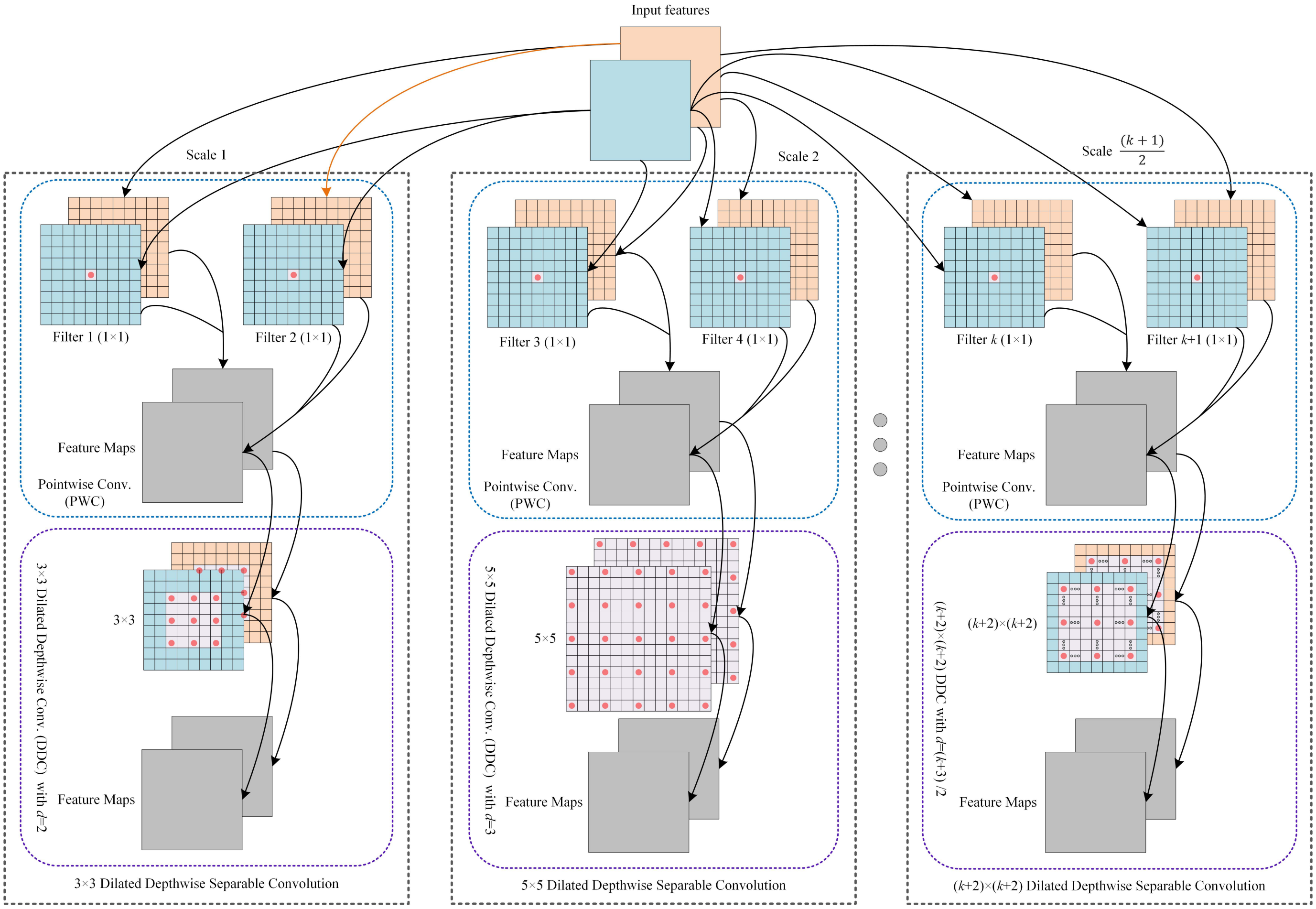
Input features
Scale 1
Scale 2
Scale $\frac{(k+1)}{2}$
Filter 1 (1×1)
Filter 2 (1×1)
Filter 3 (1×1)
Filter 4 (1×1)
Filter k (1×1)
Filter k+1 (1×1)
Feature Maps
Pointwise Conv. (PWC)
3×3 Dilated Depthwise Conv. (DDC) with d=2
3×3
5×5 Dilated Depthwise Conv. (DDC) with d=3
5×5
(k+2)×(k+2) DDC with d=(k+3)/2
(k+2)×(k+2)
3×3 Dilated Depthwise Separable Convolution
5×5 Dilated Depthwise Separable Convolution
(k+2)×(k+2) Dilated Depthwise Separable Convolution

Fig. 3: Proposed multi-scale dilated depthwise separable convolution (MDDC).

these features. Let H_f be the output features of AFP, based on Equations (2) to (4), the AFP is defined as

$$\begin{aligned} H_f &= \mathrm{AFP}(H_1; H_{\mathrm{MLKDC}}; H_{\mathrm{MDDC}}) \\ &= \frac{1}{N}\left(H_1 + \tilde{H}_3 + \tilde{H}_5 + \cdots + \tilde{H}_{m+2} + \hat{H}_3 + \hat{H}_5 + \cdots + \hat{H}_{k+2}\right), \end{aligned} \tag{5}$$

where $N = 2 + \frac{m+k}{2}$ denotes the total number of scales in MSLKC.

Analysis of Parameters and Complexity: Let the number of input channels and output channels in the ordinary convolution with a 3×3 kernel be c_i and c_o, respectively. Then the ordinary convolution has $c_i \times 3 \times 3 \times c_o$ parameters and proportional floating point operations per second (FLOPS). Assuming that DSC in Fig. 2 (b) and MSLKC have the same number of input and output channels as the ordinary convolution. Then, the DSC has $c_i \times 3 \times 3 + c_i \times c_o$ parameters and proportional FLOPS. The number of parameters for the proposed MLKDC (Fig. 2 (c)), MDDC (Fig. 3), and MSLKC modules are $\frac{m+1}{2} \times c_i \times c_o + c_o \times \sum_{j=1}^{\frac{m+1}{2}} (2j+1)^2$, $\frac{k+1}{2} \times c_i \times c_o + c_o \times \sum_{i=1}^{\frac{k+1}{2}} (2i+1)^2$, and $c_i \times c_o + \frac{m+1}{2} \times c_i \times c_o + c_o \times \sum_{j=1}^{\frac{m+1}{2}} (2j+1)^2 + \frac{k+1}{2} \times c_i \times c_o + c_o \times \sum_{i=1}^{\frac{k+1}{2}} (2i+1)^2$, respectively. The computational complexity of the three proposed modules is proportional to their respective parameters. In our experiments, we set $m = 13$, $k = 3$, and $c_i = c_o = 64$. With these settings, the receptive field of the proposed MSLKC is equivalent to that of the 7-layer 3×3 ordinary convolution (stacking the ordinary convolution seven times). The number of parameters of MSLKC and the 7-layer ordinary convolution are 86,592 and 258,048, respectively. The computational complexity of these two models is proportional to the number of parameters in their respective models. Therefore, with the same receptive field, our MSLKC outperforms multiple stacked convolutions in terms of parameters and computational complexity.

C. Softmax Classification

After AFP, we use a Softmax classifier to classify the combined feature map H_f. We have

$$Y = \frac{e^{P_i H_f + b_i}}{\sum_i^C e^{P_i H_f + b_i}}, \tag{6}$$

where C denotes the number of land cover categories, and P_i and b_i are the trainable parameter and bias. To train the proposed model, we select a cross-entropy error as the loss function, namely

$$\mathcal{L} = -\sum_{Z \in O_{\text{label}}} \sum_{j=1}^{C} O_{zj} \ln Y_{zj}, \tag{7}$$

TABLE I: SUMMARY OF BOTSWANA, HOUSTON 2013, AND LONGKOU DATASETS

Dataset	Botswana				Houston 2013				LongKou			
Wavelength	0.4 um - 2.5 um				0.38 um - 1.05 um				0.4 um - 1.0 um			
Data Size	1476 × 256 × 145				349 × 1905 × 144				550 × 400 × 270			
Ratio Per Class	1%, 1%, and 98% for Train., Val., and Test.				0.5%, 1%, and 98.5% for Train., Val., and Test.				0.025%, 1%, and 98.975% for Train., Val., and Test.			
Time	2001				2013				2018			
Class No.	Class Name	Train.	Val.	Test.	Class Name	Train.	Val.	Test.	Class Name	Train.	Val.	Test.
1	Water	3	3	264	Healthy Grass	6	13	1232	Corn	8	345	34158
2	Hippo Grass	1	1	99	Stressed Grass	6	13	1235	Cotton	2	84	8288
3	Floodplain Grasses 1	2	2	247	Synthetic Grass	4	7	686	Sesame	1	30	3000
4	Floodplain Grasses 2	2	2	211	Tree	6	12	1226	Broad-Leaf Soybean	16	632	62564
5	Reeds	3	3	263	Soil	6	12	1224	Narrow-Leaf Soybean	1	42	4108
6	Riparian	3	3	263	Water	2	3	320	Rice	3	119	11732
7	Fires Car	2	2	255	Residential	7	13	1248	Water	17	670	66369
8	Island Interior	2	2	199	Commercial	6	12	1226	Roads and Houses	2	71	7051
9	Acacia Woodlands	3	3	308	Road	6	13	1233	Mixed Weed	1	52	5176
10	Acacia Shrub Lands	2	2	244	Highway	6	12	1209	-	-	-	-
11	Acacia Grasslands	3	3	299	Railway	6	12	1217	-	-	-	-
12	Short Mopane	2	2	177	Parking Lot 1	6	12	1215	-	-	-	-
13	Mixed Mopane	3	3	262	Parking Lot 2	3	5	461	-	-	-	-
14	Exposes Soils	1	1	93	Tennis Court	2	4	422	-	-	-	-
15	-	-	-	-	Running Track	3	7	650	-	-	-	-
Total	-	32	32	3184	-	75	150	14804	-	51	2045	202446

where O denotes the label matrix, and Y_{zj} represents the probability of the z-th pixel belonging to the j-th category.

III. EXPERIMENT

In this section, we conduct comprehensive experiments to assess the strengths and effectiveness of our MSLKCNN. Firstly, we benchmark MSLKCNN against nine HSI classification methods, employing three widely utilized HSI datasets, with metrics including per-class accuracy, overall accuracy (OA), average accuracy (AA), and kappa coefficient (KAPPA). Subsequently, we compare MSLKCNN with different comparative methods under various training samples. Then, we analyze OA results using diverse fusion schemes. Furthermore, we compare the training and testing time of different methods to demonstrate the efficiency of MSLKCNN. Lastly, we perform several ablation studies to validate the impacts of key components and hyperparameters.

A. Dataset

In this section, we introduce three publicly available benchmark HSI datasets, i.e., Botswana, Houston 2013, and WHU-Hi-LongKou (LongKou), to evaluate the performance of the proposed MSLKCNN. The details of the three datasets are summarized in Table I.

1) Botswana: The first dataset, Botswana, was acquired by the NASA EO-1 satellite in the Okavango Delta region of Botswana. It contains 242 spectral bands with a spatial size of 1476×256 and 14 land cover categories in the wavelength range from 0.4 to 2.5 um. After removing noise bands of 1-9, 56-81, 98-101,120-133, and 165-186, 145 spectral bands are retained. Moreover, 1%, 1%, and 98% of samples per class are randomly selected for training, validation, and testing, respectively.

2) Houston 2013: The second dataset, Houston 2013, was collected by the National Center for Airborne Laser Mapping (NCALM) over the University of Houston for the 2013 IEEE GRSS Data Fusion Contest [77]. The dataset contains 144 spectral bands with a spatial size of 349×1905 and 15 land cover categories ranging from 0.38 to 1.05 um. In addition, we randomly select 0.5%, 1%, and 98.5% of samples per class for training, validation, and testing, respectively.

3) WHU-Hi-LongKou: The third dataset, WHU-Hi-LongKou (LongKou), was captured by an 8-mm focal length Headwall Nano-Hyperspec imaging sensor in the town of LongKou, Hubei Province, China in 2018 [78]. We use the image with a spatial size of 550×400, 9 land cover classes, and 240 spectral bands in the wavelength range from 0.4 to 1.0 um. Then, 0.025%, 1%, and 98.975% of samples per class are randomly selected as the training, validation, and testing sets, respectively.

B. Experimental Settings

We conduct all experiments using Adam optimizer with a learning rate of 0.0005 in Pytorch. In NSM, we set the number of filters per convolutional layer to 64. For the proposed MSLKC, the number of filters per convolution is set to 64, the largest kernel size $(m+2)$ in MLKDC is 15 (i.e., $m = 13$), and the largest kernel size $(k+2)$ in MDDC is 5 (i.e., $k = 3$). The number of training epochs is set to 200 for Houston 2013 and 800 for other datasets, respectively. All experiments of our MSLKCNN and baselines are repeated ten times with varying random initializations. The experimental environment consists of an i9-7980XE CPU, Python 3.7, and a GTX-2080Ti GPU.

To demonstrate the proposed MSLKCNN, we compare it against nine state-of-the-art HSI classification baselines: (1) three CNN-based methods: the deeper with contextual CNN (CDCNN) [40], the attention-based adaptive spectral-spatial kernel ResNet (A^2S^2K-Res) [79], and the central vector oriented self-similarity network (CVSSN) [80]; (2) three GCN-based methods: the CNN-enhanced GCN (CEGCN) [55], the fast dynamic graph convolutional network and CNN parallel network (FDGC) [81], and the attention multi-hop graph and multiscale convolutional fusion network (AMGCFN) [82]; and (3) three Transformer-based methods: the spectral-spatial feature tokenization transformer (SSFTT) [7], the GCN and transformer fusion network (GTFN) [65], and the groupwise separable convolutional vision Transformer (GSC-ViT) [66].

TABLE II: COMPARISON OF ALL METHODS ON THE BOTSWANA DATASET USING 1% LABELLED SAMPLES PER CLASS FOR TRAINING. LKCNet DENOTES LARGE KERNEL CONVOLUTIONAL NETWORK.

Class	CNN-based methods			GCN-based methods			Transformer-based methods			LKCNet
	CDCNN	A^2S^2K-Res	CVSSN	CEGCN	FDGC	AMGCFN	SSFTT	GTFN	GSC-ViT	MSLKCNN
1	**100.0±0.00**	77.35±14.69	90.36±3.82	98.79±1.65	75.68±14.09	98.48±1.93	96.74±2.29	74.46±11.77	98.56±0.77	**100.0±0.00**
2	17.98±12.51	74.55±20.14	79.08±15.22	87.07±19.43	85.45±12.20	97.37±4.76	**100.0±0.00**	74.40±15.78	99.19±1.62	94.65±7.17
3	35.06±37.07	83.00±15.01	91.66±9.93	84.13±10.24	82.12±10.25	94.49±3.67	78.62±8.50	79.92±11.74	80.65±11.72	**95.10±4.90**
4	10.24±18.14	99.62±0.76	75.47±11.71	87.58±16.78	73.65±24.33	87.77±14.43	**100.0±0.00**	88.08±18.40	96.59±6.36	**100.0±0.00**
5	67.83±8.28	84.41±6.90	77.02±6.23	56.35±18.50	77.26±13.92	86.24±4.03	77.87±6.23	71.65±7.31	84.71±10.37	**95.74±1.49**
6	45.93±22.56	73.54±17.97	74.38±11.97	72.02±6.26	65.93±11.13	67.30±0.76	55.89±0.42	69.70±5.35	**79.92±5.19**	67.91±6.84
7	88.47±5.43	94.35±5.07	95.20±3.56	95.14±6.23	93.44±8.04	99.61±0.50	**99.84±0.31**	94.22±6.49	**99.84±0.31**	98.82±2.90
8	24.62±23.16	68.04±15.52	74.24±14.04	65.93±5.61	57.79±16.79	66.73±4.49	**97.99±2.93**	60.00±7.45	96.28±2.67	71.36±0.00
9	71.49±17.68	93.83±6.43	68.80±14.89	**100.0±0.00**	87.01±15.68	**100.0±0.00**	78.38±7.63	71.58±15.89	89.55±4.87	**100.0±0.00**
10	47.30±34.75	75.74±25.30	92.97±5.67	**100.0±0.00**	69.43±15.29	91.15±14.97	**100.0±0.00**	62.03±21.85	98.69±2.03	93.81±10.09
11	98.06±3.71	84.41±12.77	88.54±9.17	98.13±2.14	79.80±29.80	96.99±3.42	76.52±0.54	95.23±4.59	89.77±4.21	**99.13±1.11**
12	90.40±8.33	81.36±21.52	66.24±6.97	95.59±7.99	94.12±11.19	**100.0±0.00**	70.51±0.23	75.87±25.01	99.66±0.45	**100.0±0.00**
13	80.84±22.70	92.75±10.81	90.57±8.49	**100.0±0.00**	92.67±13.15	**100.0±0.00**	98.78±1.40	89.06±8.28	99.62±0.76	**100.0±0.00**
14	23.87±29.00	65.16±17.70	**97.93±4.14**	48.39±21.71	56.77±19.68	16.13±0.00	50.97±18.76	29.57±14.55	48.39±19.32	90.75±10.77
OA	62.39±9.25	83.49±1.94	81.71±2.1	87.08±1.15	78.84±3.89	89.20±1.43	84.99±0.78	76.52±2.50	91.33±1.75	**93.74±1.23**
AA	57.29±9.72	82.01±2.38	83.03±1.39	84.94±0.92	77.94±4.07	85.88±1.59	84.44±1.04	73.98±2.10	90.10±2.20	**93.38±1.40**
KAPPA	59.03±10.11	82.09±2.11	80.17±2.29	85.97±1.25	77.06±4.22	88.27±1.56	83.74±0.84	74.50±2.71	90.61±1.90	**93.21±1.34**

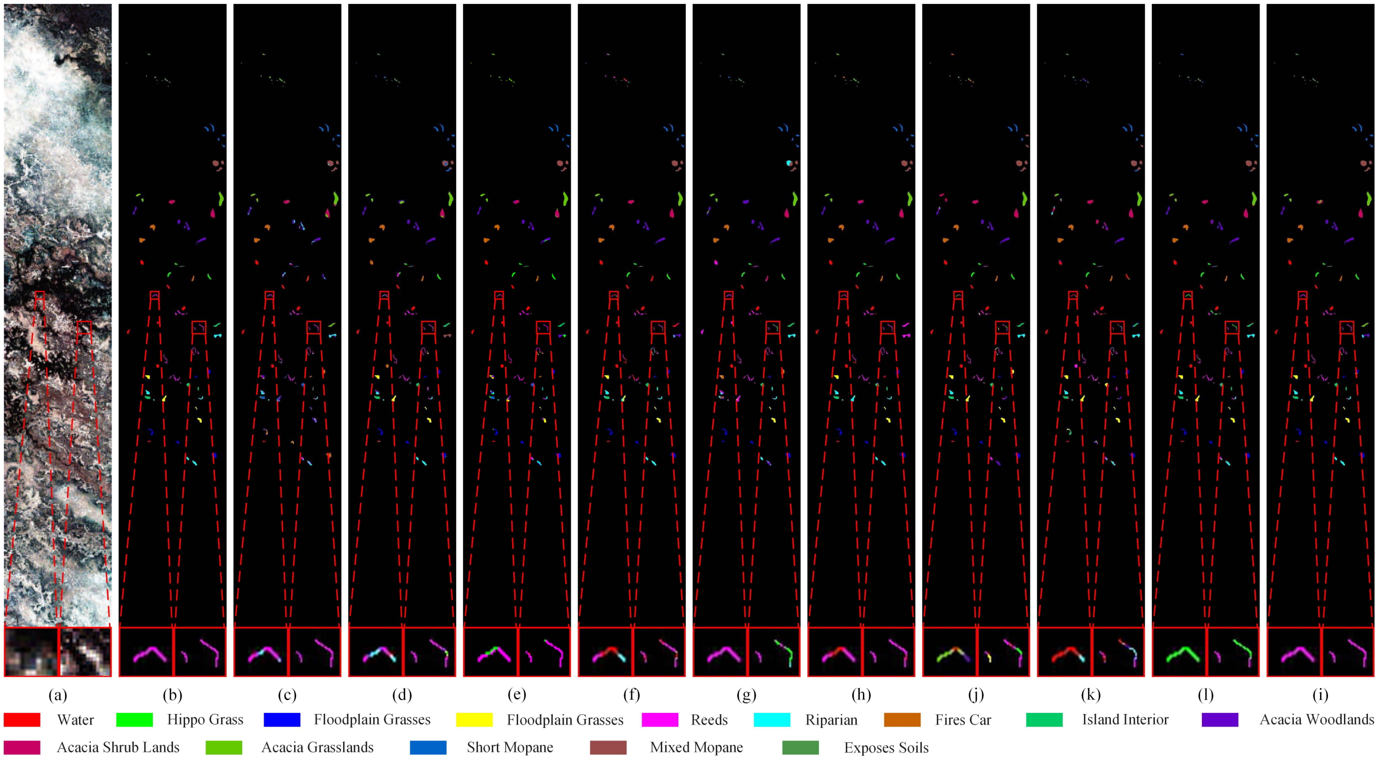

Fig. 4: False-color image, ground truth, and classification maps on the Botswana dataset. (a) False-color image. (b) Ground truth. (c) CDCNN (OA=62.39%). (d) A^2S^2K-Res (OA=83.49%). (e) CVSSN (OA=81.71%). (f) CEGCN (OA=87.08%). (g) FDGC (OA=78.84%). (h) AMGCFN (OA=89.20%). (j) SSFTT (OA=84.99%). (k) GTFN (OA=76.52%). (l) GSC-ViT (OA=91.33%). (i) MSLKCNN (OA=93.74%).

C. Comparison of Classification Performance

In this section, we assess the performance of the proposed MSLKCNN quantitatively and qualitatively against the baselines on the Botswana, Houston 2013, and LongKou datasets.

1) Results on Botswana: The quantitative results of all methods on the Botswana dataset are reported in Table II. From the table, we make the following observations: a) Compared to CDCNN that lacks these techniques such as residual connections, attention mechanisms, and multi-branch learning, other baselines that incorporate at least one of these techniques perform significantly better in terms of OA, AA, and KAPPA. This demonstrates that these techniques can enhance the performance of HSI classification. b) Most GCN-based (e.g., AMGCFN) and Transformer-based (e.g., GSC-ViT) methods outperform CNN-based methods. This superiority is primarily attributed to the ability of GCNs and Transformers to establish long-range dependencies among pixels. c) By leveraging large kernel convolutions, our MSLKCNN is capable of capturing global features that are often overlooked by traditional CNNs, achieving top-level performance among all competing methods in terms of OA, AA, and KAPPA. This validates the effectiveness of MSLKCNN. The classification maps of different methods on the dataset are shown in Fig. 4.

TABLE III: COMPARISON OF ALL METHODS ON THE HOUSTON 2013 DATASET USING 0.5% LABELLED SAMPLES PER CLASS FOR TRAINING. LKCNet DENOTES LARGE KERNEL CONVOLUTIONAL NETWORK.

Class	CNN-based methods			GCN-based methods			Transformer-based methods			LKCNet
	CDCNN	A^2S^2K-Res	CVSSN	CEGCN	FDGC	AMGCFN	SSFTT	GTFN	GSC-ViT	MSLKCNN
1	84.24±1.25	88.81±6.87	85.77±8.70	93.93±2.97	86.01±5.28	90.89±2.76	91.61±1.80	84.50±1.96	**98.23±0.52**	96.69±2.30
2	34.56±21.86	**94.95±6.11**	84.47±8.11	81.31±5.55	74.57±10.31	68.70±5.18	83.53±9.79	82.60±7.18	81.28±6.42	85.78±0.79
3	86.65±4.58	99.77±0.07	77.49±10.96	**99.91±0.12**	95.78±3.20	99.10±0.63	97.70±1.46	98.90±0.67	95.19±1.99	98.54±0.60
4	91.21±6.74	**95.20±3.08**	89.57±5.38	92.95±1.68	72.04±6.29	77.59±4.43	88.52±7.21	85.98±8.88	92.92±1.95	94.13±0.25
5	93.22±3.63	99.22±0.97	85.87±6.63	**100.0±0.00**	98.24±1.99	97.79±1.19	99.40±0.58	97.17±2.19	99.31±0.40	99.40±0.51
6	32.81±12.87	72.69±31.04	83.43±14.91	79.00±3.08	80.50±5.24	63.13±16.47	88.00±5.97	81.61±6.92	**89.06±3.94**	79.25±2.51
7	61.47±13.52	85.35±4.20	76.19±6.16	85.16±3.81	53.56±11.57	65.00±2.49	65.77±10.63	58.91±9.78	**86.51±5.95**	83.29±3.16
8	33.00±3.82	52.06±4.89	**76.31±6.21**	48.43±1.91	48.76±9.21	40.28±4.62	48.65±10.33	50.55±11.85	51.24±2.36	46.46±8.19
9	65.37±9.43	63.29±9.70	79.56±7.40	**86.36±4.17**	51.71±15.97	56.25±7.31	49.46±8.95	57.35±8.91	76.66±2.75	85.52±3.92
10	21.59±11.08	68.90±10.09	57.66±7.56	73.27±0.40	**81.60±10.15**	73.10±0.43	56.29±7.41	68.52±14.16	62.03±0.52	71.99±1.11
11	35.76±19.15	84.08±3.51	59.88±11.49	82.09±1.31	79.90±8.63	78.72±4.19	67.54±3.51	84.07±11.13	63.16±8.09	**87.90±1.33**
12	32.02±20.83	74.06±10.24	65.25±10.62	**87.16±4.71**	77.79±10.63	79.72±12.76	75.98±4.10	62.92±18.51	50.72±5.05	73.33±8.37
13	64.77±14.69	80.56±9.47	75.34±15.02	38.18±18.59	66.41±20.42	**88.72±3.13**	76.75±14.30	52.55±12.99	78.96±5.75	76.10±8.21
14	51.04±28.33	98.96±1.75	68.76±7.14	98.77±1.19	97.91±2.53	93.89±4.88	99.95±0.09	96.95±1.84	91.52±6.68	**100.0±0.00**
15	89.82±8.10	**99.94±0.12**	78.07±10.62	98.43±0.81	92.89±4.65	94.34±4.74	93.97±7.04	98.78±1.33	99.60±0.49	99.91±0.12
OA	57.98±3.55	82.71±0.66	74.83±1.77	83.50±0.42	75.06±2.92	75.83±2.69	76.01±2.67	75.76±1.45	78.98±1.01	**84.22±1.24**
AA	58.50±3.57	83.86±1.29	76.24±1.26	83.00±1.04	77.18±3.30	77.81±2.82	78.87±2.32	77.42±1.18	81.09±0.75	**85.22±1.21**
KAPPA	54.68±3.85	81.31±0.72	72.80±1.93	82.14±0.46	73.05±3.16	73.91±2.88	74.08±2.87	73.79±1.56	77.27±1.09	**82.93±1.34**

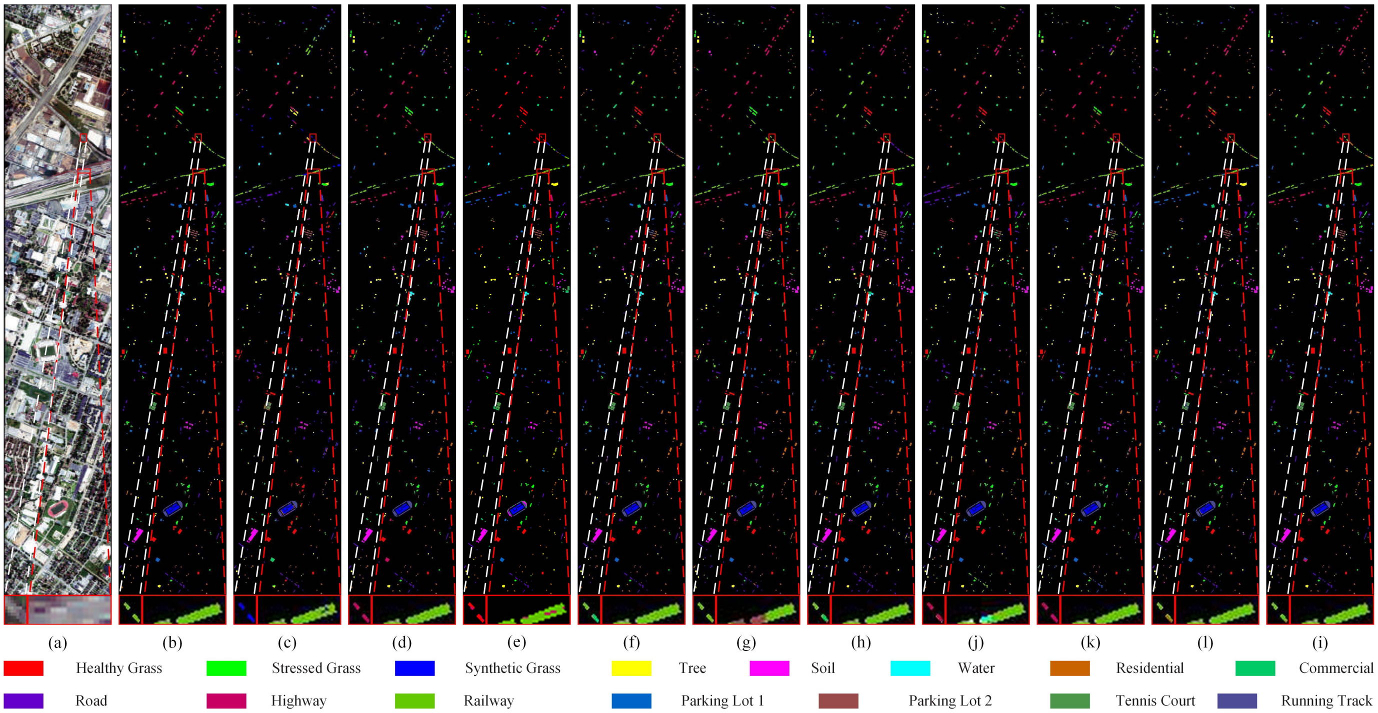

Fig. 5: False-color image, ground truth, and classification maps on the Houston 2013 dataset. (a) False-color image. (b) Ground truth. (c) CDCNN (OA=57.98%). (d) A^2S^2K-Res (OA=82.71%). (e) CVSSN (OA=74.83%). (f) CEGCN (OA=83.50%). (g) FDGC (OA=75.06%). (h) AMGCFN (OA=75.83%). (j) SSFTT (OA=76.01%). (k) GTFN (OA=75.76%). (l) GSC-ViT (OA=78.98%). (i) MSLKCNN (OA=84.22%).

From these maps, we observe that the proposed MSLKCNN achieves fewer misclassifications than the comparison methods.

2) Results on Houston 2013: Table III presents the quantitative results of the proposed MSLKCNN in comparison with the state-of-the-art methods on the Houston 2013 dataset. From these results, it is evident that our MSLKCNN achieves the best performance, surpassing all other methods in terms of OA, AA, and KAPPA. Specifically, MSLKCNN improves over CNN-based methods by at least 1.83%, 1.62%, and 1.99%, improves over GCN-based methods by at least 0.86%, 2.67%, and 0.96%, and improves over Transformer-based methods by at least 6.63%, 5.09%, and 7.32% in terms of OA, AA, and KAPPA, respectively. These results demonstrate the strengths of the proposed MSLKCNN. As shown in Fig. 5, MSLKCNN achieves superior classification map compared to other methods, which further validates that our large kernel convolution contributes significantly to enhancing performance.

3) Results on LongKou 2013: Table IV shows the quantitative results achieved by various methods on the LongKou dataset. Similar to these observations for other datasets, the proposed MSLKCNN exhibits a substantial improvement over all baselines, again demonstrating the superiority of MSLKCNN. The visual inspection in Fig. 6 reveals that MSLKCNN achieves fewer misclassifications than the comparison methods, such as the class of Roads and Houses (masked in green).

TABLE IV: COMPARISON OF ALL METHODS ON THE LONGKOU DATASET USING 0.025% LABELLED SAMPLES PER CLASS FOR TRAINING. LKCNet DENOTES LARGE KERNEL CONVOLUTIONAL NETWORK.

Class	CNN-based methods			GCN-based methods			Transformer-based methods			LKCNet
	CDCNN	A^2S^2K-Res	CVSSN	CEGCN	FDGC	AMGCFN	SSFTT	GTFN	GSC-ViT	MSLKCNN
1	96.48±1.15	**99.89±0.08**	88.01±6.81	99.07±0.33	97.88±2.17	99.35±0.28	98.55±0.92	97.44±1.91	99.11±0.49	99.58±0.32
2	54.13±31.60	64.04±23.11	55.72±8.34	52.25±1.90	81.38±14.11	72.93±6.84	70.74±13.34	60.67±19.75	**86.19±7.09**	65.03±3.74
3	0.51±1.01	66.84±34.34	45.63±38.32	51.81±15.88	88.67±6.57	92.93±1.43	**95.71±4.54**	63.94±30.20	82.63±4.66	81.06±1.20
4	91.14±3.77	98.62±0.75	91.05±2.15	**99.49±0.23**	96.70±2.62	96.60±0.34	94.06±3.43	94.23±6.58	97.45±0.95	99.45±0.16
5	0.01±0.02	33.95±23.10	40.88±28.90	28.04±3.84	**59.44±8.37**	56.58±13.33	30.96±3.21	38.73±14.96	22.30±5.76	54.69±12.05
6	80.36±4.38	82.11±7.26	83.27±5.45	93.56±2.74	88.20±2.87	84.17±4.32	**97.55±1.13**	85.42±7.96	77.38±7.05	97.00±1.43
7	**99.99±0.00**	99.86±0.15	98.94±0.66	**99.99±0.01**	98.43±1.69	99.81±0.11	97.95±0.58	98.22±1.12	99.88±0.14	**99.99±0.01**
8	52.71±35.51	47.08±18.07	61.31±21.69	43.63±13.61	39.55±12.82	68.28±5.23	67.17±3.30	54.63±22.72	83.36±6.50	**88.37±4.78**
9	1.49±2.98	35.11±8.98	**71.08±31.19**	2.45±1.46	27.48±16.13	39.51±20.19	21.22±4.99	24.86±14.02	19.54±5.70	21.86±6.89
OA	85.98±1.36	91.66±1.64	87.92±0.84	90.72±0.60	91.71±0.53	93.12±0.39	91.29±1.28	89.47±1.13	92.68±0.44	**94.55±0.36**
bAA	52.98±2.84	69.72±7.34	70.65±5.13	63.37±2.29	75.30±3.52	78.91±1.56	74.88±1.61	68.68±9.28	74.21±1.37	**78.56±1.92**
KAPPA	81.35±1.78	88.88±2.22	84.01±1.11	87.45±0.84	89.00±69.07	90.87±0.52	88.50±1.69	85.97±1.64	90.29±0.60	**92.76±0.48**

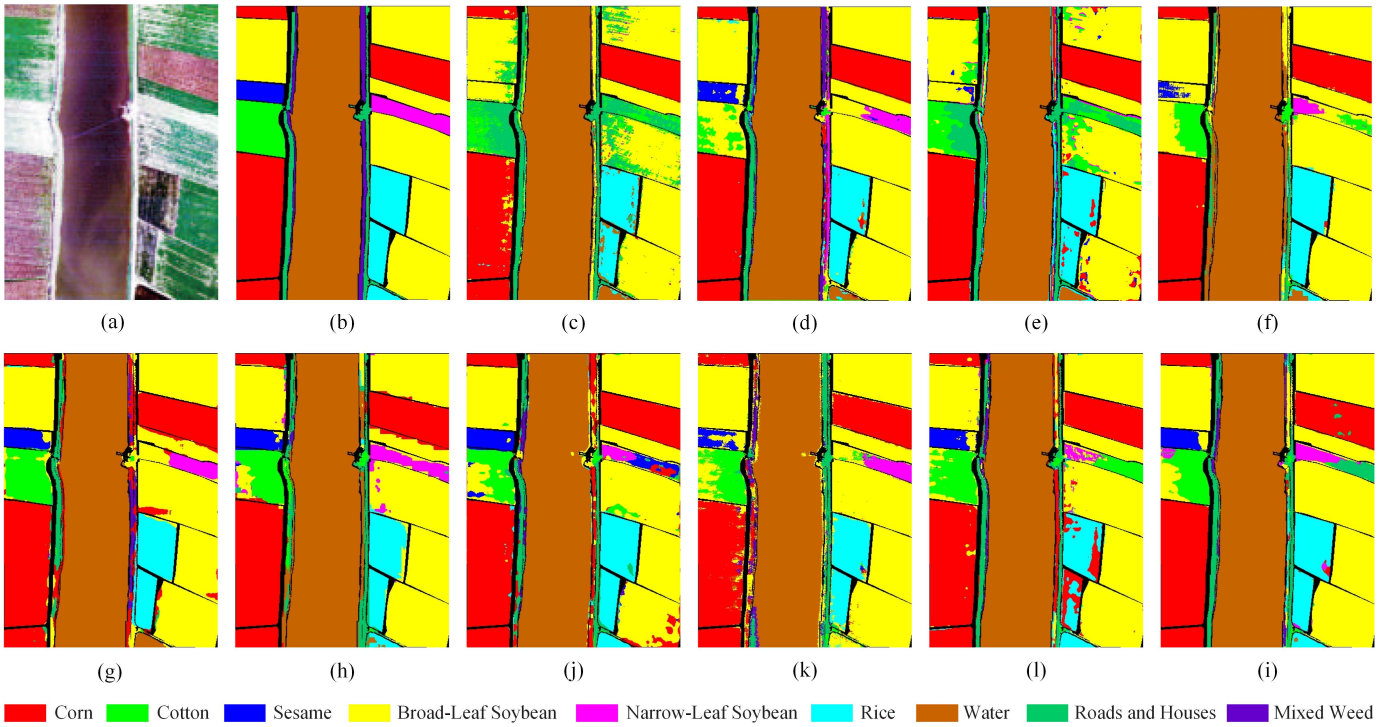

Fig. 6: False-color image, ground truth, and classification maps on the LongKou dataset. (a) False-color image. (b) Ground truth. (c) CDCNN (OA=85.98%). (d) A^2S^2K-Res (OA=91.66%). (e) CVSSN (OA=87.92%). (f) CEGCN (OA=90.72%). (g) FDGC (OA=91.71%). (h) AMGCFN (OA=93.12%). (j) SSFTT (OA=91.29%). (k) GTFN (OA=89.47%). (l) GSC-ViT (OA=92.68%). (i) MSLKCNN (OA=94.55%).

D. Analysis of All methods with Different Numbers of Training Samples

In this section, we compare the OA performance of various methods under varying percentages of training samples per class, i.e., 0.5%, 1%, 2.5%, 5%, and 10% for the Botswana dataset, 0.125%, 0.25%, 0.5%, 1%, and 1.5% for the Houston 2013 dataset, and 0.01%, 0.025%, 0.05%, 0.1%, and 0.2% for the LongKou dataset. The percentage of validation samples is set to 1% for these methods across each dataset. As depicted in Fig. 7, the OA results of the ten methods typically improve with an increase in the percentage of training samples. We observe that in a few cases, the OA outcomes with more training samples decrease for several comparative methods (such as GSC-ViT). These abnormal results may be due to the additional noise introduced by the increase in the number of training samples. In addition, the OA results of CEGCN, AMGCFN, and the proposed MSLKCNN show an improvement as the number of training samples increases, which is attributed to the noise suppression module in their models. Furthermore, our MSLKCNN outperforms the comparison methods across different datasets, further demonstrating the strengths of MSLKCNN.

E. Analysis of Different Fusion Schemes

As described in Section II-B, the number of feature maps of the proposed MLKDC and MDDC is substantial. Using column concatenation fusion (concatenate) and sum fusion (sum) to combine these features may respectively lead to an increase in the number of parameters and the generation of large feature values. This may potentially lead to overfitting and gradient explosion problems, respectively. To avoid these potential problems, we introduce the AFP fusion scheme. To

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3566616

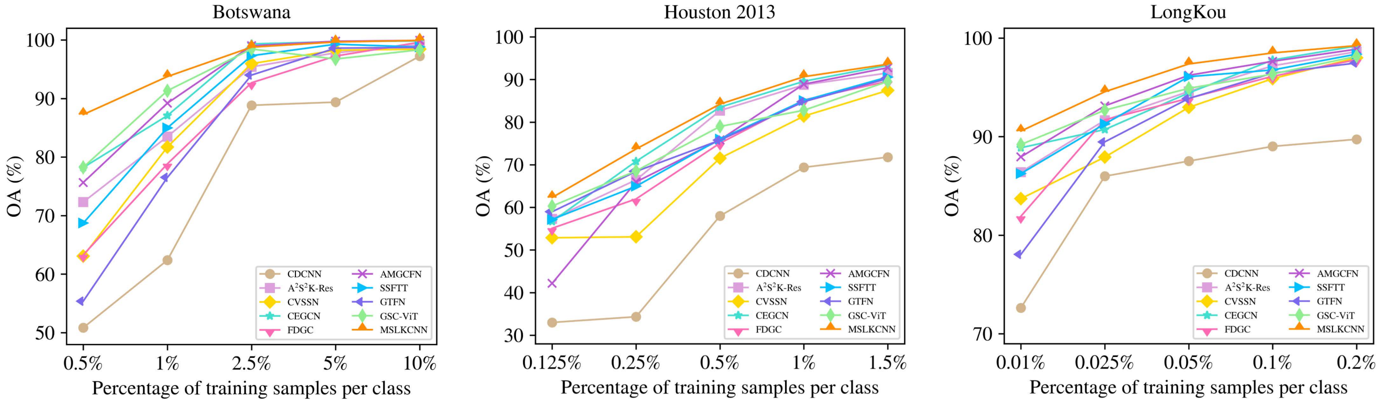

Fig. 7: OA performance of diverse methods under varying percentages of training samples per class for the Botswana, Houston 2013, and LongKou datasets.

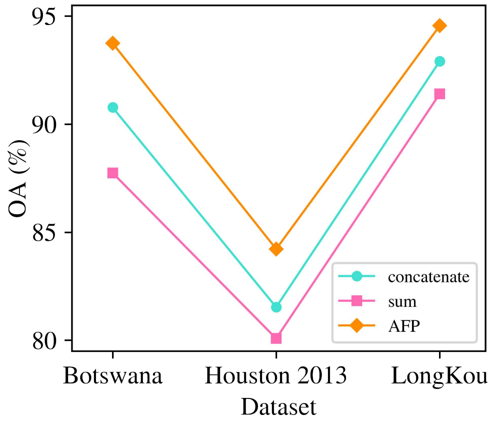

Fig. 8: Results of various fusion schemes on the Botswana, Houston 2013, and LongKou datasets.

assess the proposed AFP, we compare the OA of AFP with the two commonly fusion schemes, i.e., concatenate and sum. The results are illustrated in Fig. 8. The results show that AFP performs better OA compared to other fusion schemes, which verifies the advantages of designing AFP.

F. Comparison of Running Time

Table V reports the results of diverse methods in both training and testing time on GPU and CPU devices for all datasets. From the results, we make the following observations: 1) CDCNN, CVSSN, and SSFTT perform faster in most cases compared to other baselines. This is attributed to the fact that the three models incorporate a small number of convolutional layers with kernel sizes larger than 1×1. 2) CEGCN and AMGCFN, which incorporate GCN modules, utilize the entire HSI instead of small HSI cubes as input, which results in faster prediction speed than most comparative methods. However, when dealing with the large datasets such as Botswana and Houston 2013, they require considerable memory resources, leading to the memory overflow on GPU. 3) Similar to CEGCN and AMGCFN, the proposed MSLKCNN also leverages the whole HSI as input, achieving very short prediction time on GPU for all datasets, especially on the LongKou dataset. In addition, MSLKCNN outperforms all compared methods by a significant margin. These observations demonstrate the superiority of MSLKCNN with large kernel convolution for applications in the industrial world.

G. Ablation Study

In this section, we conduct ablation experiments to verify the contributions and effects of various components and the two key hyperparameters within the proposed MSLKCNN: different largest kernel sizes in MLKDC and diverse largest kernel sizes in MDDC.

1) Contributions of Various Components: As shown in Fig. 1, our MSLKCNN is mainly composed of four components, i.e., the NSM, the 1×1 convolution block, MLKDC, and MDDC. To evaluate the individual contributions of these components, we make a quantitative comparison by removing one of the four components. The results are summarized in Table VI. We observe that MSLKCNN without NSM performs significantly lower than other methods on Houston 2013 dataset, suggesting that the dataset may contain considerable noise. Additionally, we find that MSLKCNN without MLKDC performs worse compared to other models on other datasets, which demonstrates that the large kernel convolutions in MLKDC enhance the capability of extracting features. Furthermore, we see that our MSLKCNN outperforms each modified method of MSLKCNN across all datasets. These results validate the effectiveness of these components.

2) Analysis of Different Largest Kernel Sizes in MLKDC: To investigate the impact of varying the number of largest kernel sizes in MLKDC, we compare the OA results achieved with different largest kernel sizes (in MLKDC) on the Botswana, Houston 2013, and LongKou datasets. Fig. 9 presents the quantitative results. We observe that, in MSLKCNN, the OA generally improves as the kernel size increases until its value of 15×15. By further increasing the value, however, the OA begins to decrease. This is instrumental in determining the optimal largest kernel size for MLKDC. The optimal largest kernel size of MLKDC is the best largest kernel size for MSLKC.

3) Analysis of Diverse Largest Kernel Sizes in MDDC: To verify the impact of varying the number of largest kernel sizes in MDDC, we compare the OA results using diverse largest

TABLE V: RUNNING TIME OF VARIOUS METHODS ON THE BOTSWANA, HOUSTON 2013, AND LONGKOU DATASETS. OOM: OUT OF MEMORY. LKCNet DENOTES LARGE KERNEL CONVOLUTIONAL NETWORK. ms DENOTES MILLISECOND.

Dataset	Time (second)	CNN-based methods			GCN-based methods			Transformer-based methods			LKCNet
		CDCNN	A^2S^2K-Res	CVSSN	CEGCN	FDGC	AMGCFN	SSFTT	GTFN	GSC-ViT	MSLKCNN
Botswana	Train. (GPU)	2.95	5.03	5.06	OOM	4.89	OOM	**1.82**	30.73	11.12	337.47
	Test. (GPU)	3.44	0.82	**7.72 ms**	OOM	0.33	OOM	0.18	3.65	49.52	0.02
	Train. (CPU)	5.48	10.52	16.12	1246.43	9.93	6584.85	**4.20**	48.69	30.38	8037.40
	Test. (CPU)	1.73	2.84	**0.01**	0.75	1.54	4.60	0.99	7.80	381.39	3.09
Houston 2013	Train. (GPU)	8.12	8.99	8.74	OOM	5.24	OOM	**4.37**	64.40	17.96	439.14
	Test. (GPU)	3.42	4.07	**3.74 ms**	OOM	1.58	OOM	0.83	1.57	79.82	1.19
	Train. (CPU)	**10.52**	17.97	37.90	1597.37	14.04	2493.61	12.11	107.66	71.89	4375.00
	Test. (CPU)	7.74	13.36	**0.02**	5.71	7.20	9.01	4.67	35.76	630.68	8.92
LongKou	Train. (GPU)	23.63	38.53	OOM	360.58	**4.44**	186.22	13.72	OOM	169.64	204.02
	Test. (GPU)	47.79	50.13	OOM	0.29	17.67	0.04	10.30	OOM	46.60	**0.02**
	Train. (CPU)	59.51	121.51	277.66	1726.91	**10.87**	3810.58	70.30	67.70	507.01	4632.91
	Test. (CPU)	142.63	188.65	**58.99 ms**	1.27	90.51	2.67	63.49	497.74	288.18	1.89

TABLE VI: CLASSIFICATION RESULTS OF EACH COMPONENT IN MSLKCNN

NSM	1 × 1 Conv. block	MLKDC	MDDC	Botswana			Houston 2013			LongKou		
				OA	AA	KAPPA	OA	AA	KAPPA	OA	AA	KAPPA
✕	✓	✓	✓	87.54	84.41	86.46	78.40	79.14	76.63	94.17	78.15	92.23
✓	✕	✓	✓	93.46	92.90	92.91	83.20	83.98	81.82	94.09	76.49	92.14
✓	✓	✕	✓	83.91	82.54	82.54	82.52	82.43	81.08	89.49	66.44	85.97
✓	✓	✓	✕	92.92	92.54	92.32	83.47	84.64	82.13	94.27	77.23	92.38
✓	✓	✓	✓	**93.74**	**93.38**	**93.21**	**84.22**	**85.22**	**82.93**	**94.55**	**78.56**	**92.76**

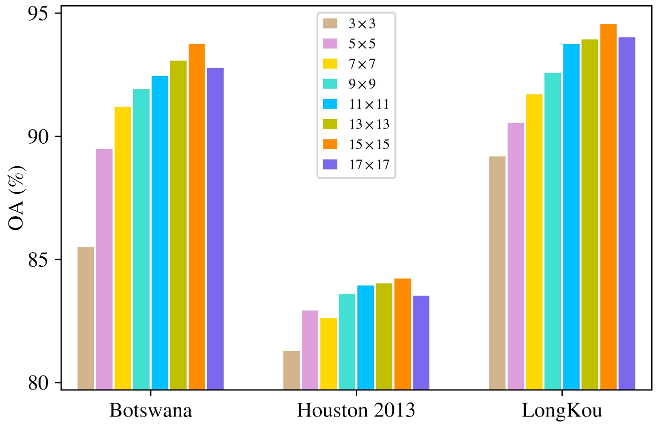

Fig. 9: Results of different largest kernel sizes in MLKDC on the Botswana, Houston 2013, and LongKou datasets.

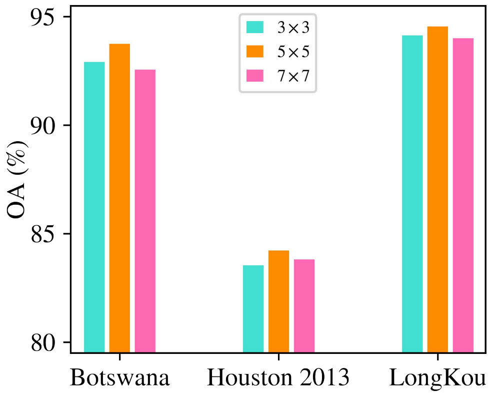

Fig. 10: Results of diverse largest kernel sizes in MDDC on the Botswana, Houston 2013, and LongKou datasets.

kernel sizes (in MDDC) on the Botswana, Houston 2013, and LongKou datasets. The quantitative results are reported in Fig. 10. We find that, in MSLKCNN, the OA achieves an improvement as the kernel size increases from 3 × 3 to 5 × 5. However, further enlargement of the size leads to a decrease in OA. This finding is crucial in identifying the optimal largest kernel size for MDDC.

IV. Conclusion

In this work, we introduce a novel approach to HSI classification by designing large kernel CNN architecture. The proposed model employs NSM and MSLKC components to suppress noise and extract robust spectral-spatial features, respectively. Among these components, MSLKC, as the single-layer feature extraction layer, is the main highlight. It consists of three parallel convolution operations: 1) a 1 × 1 convolution block for extracting spectral features, 2) MLKDC that is used to capture spectral-spatial features across diverse ranges (short-range, medium-range, and long-range), and 3) MDDC designed to aggregate spectral-spatial features between land covers at various distances. Our large kernel model, with its simple structure, utilizes multi-scale large kernel structure, DSCs, and dilated convolutions to effectively extract both local and global features. Employing DSCs instead of standard convolutions significantly reduces the number of parameters and computational requirements. Consequently, the proposed MSLKCNN is adaptable to the Botswana, Houston 2013, and Longkou datasets, as well as other HSI datasets with varying spectral and spatial characteristics. Extensive experiments conducted on these three datasets demonstrate the effectiveness and strengths of the proposed model. Additionally, these advantages should facilitate its extension to other types of remote sensing data, such as multispectral and synthetic aperture radar (SAR) images, with minimal modifications. SAR data, in particular, presents unique challenges due to its sensitivity to surface texture and structure rather than spectral information, which often leads to speckle noise. The multi-scale large kernel structure in MSLKCNN could be beneficial for capturing spatial patterns at different scales, and the DSCs may help in managing computational complexity when processing SAR data. For future work, an interesting direction is to design parameter sharing among different convolutional kernels to reduce memory resources. Furthermore, we plan to expand the proposed MSLKCNN to more HSI datasets.

References

[1] M. Govender, K. Chetty, and H. Bulcock, “A review of hyperspectral remote sensing and its application in vegetation and water resource studies,” *Water Sa*, vol. 33, no. 2, pp. 145–151, Apr. 2007.

[2] D. Ramakrishnan and R. Bharti, “Hyperspectral remote sensing and geological applications,” *Current sci.*, vol. 108, no. 5, pp. 879–891, Mar. 2015.

[3] Z. Pan, G. Healey, M. Prasad, and B. Tromberg, “Face recognition in hyperspectral images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1552–1560, Dec. 2003.

[4] Y. Ding, X. Zhao, Z. Zhang, W. Cai, and N. Yang, “Graph sample and aggregate-attention network for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Mar. 2021.

[5] G. Lu and B. Fei, “Medical hyperspectral imaging: a review,” *J. Biomed. Opt.*, vol. 19, no. 1, pp. 010 901:1–010 901:23, Jan. 2014.

[6] L. Mou, P. Ghamisi, and X. X. Zhu, “Deep recurrent neural networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Apr. 2017.

[7] 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, pp. 1–14, Jan. 2022.

[8] X. Yang, Y. Ye, X. Li, R. Y. 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, Sept. 2018.

[9] Y. Ding, Z. Zhang, X. Zhao, D. Hong, W. Li, W. Cai, and Y. Zhan, “Af2gnn: Graph convolution with adaptive filters and aggregator fusion for hyperspectral image classification,” *Inf. Sci.*, vol. 602, pp. 201–219, Jul. 2022.

[10] M. Pal and P. M. Mather, “An assessment of the effectiveness of decision tree methods for land cover classification,” *Remote Sens. Environ.*, vol. 86, no. 4, pp. 554–565, Aug. 2003.

[11] E. Blanzieri and F. Melgani, “Nearest neighbor classification of remote sensing images with the maximal margin principle,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1804–1811, Jun. 2008.

[12] H. Yuan and Y. Y. Tang, “Spectral–spatial shared linear regression for hyperspectral image classification,” *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 934–945, Apr. 2017.

[13] M. Belgiu and L. Drăguţ, “Random forest in remote sensing: A review of applications and future directions,” *ISPRS J. Photogramm.*, vol. 114, pp. 24–31, Apr. 2016.

[14] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, “Spectral and spatial classification of hyperspectral data using svms and morphological profiles,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[15] X. Sun, Q. Qu, N. M. Nasrabadi, and T. D. Tran, “Structured priors for sparse-representation-based hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 7, pp. 1235–1239, Jul. 2014.

[16] L. Shen and S. Jia, “Three-dimensional gabor wavelets for pixel-based hyperspectral imagery classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 5039–5046, Dec. 2011.

[17] A. Plaza, P. Martinez, R. Perez, and J. Plaza, “A new approach to mixed pixel classification of hyperspectral imagery based on extended morphological profiles,” *Pattern Recogn.*, vol. 37, no. 6, pp. 1097–1116, Jun. 2004.

[18] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “Cospace: Common subspace learning from hyperspectral-multispectral correspondences,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Jul. 2019.

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

[20] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, “Cascaded recurrent neural networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[21] W. Li, G. Wu, F. Zhang, and Q. Du, “Hyperspectral image classification using deep pixel–pair features,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[22] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, “Deep supervised learning for hyperspectral data classification through convolutional neural networks,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4959–4962.

[23] Y. Li, H. Zhang, and Q. Shen, “Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network,” *Remote Sens.*, vol. 9, no. 1, p. 67, Jan. 2017.

[24] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, “Learning and transferring deep joint spectral–spatial features for hyperspectral classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.

[25] M. Zhang, W. Li, and Q. Du, “Diverse region–based cnn for hyperspectral image classification,” *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.

[26] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. Plaza, J. Li, and F. Pla, “Capsule networks for hyperspectral image classification,” *IEEE Trans. Image Process.*, vol. 57, no. 4, pp. 2145–2160, Apr. 2019.

[27] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, “Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 241–245, Feb. 2019.

[28] A. Sha, B. Wang, X. Wu, and L. Zhang, “Semisupervised classification for hyperspectral images using graph attention networks,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 157–161, Jan. 2021.

[29] J. Bai, B. Ding, Z. Xiao, L. Jiao, H. Chen, and A. C. Regan, “Hyperspectral image classification based on deep attention graph convolutional network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Mar. 2021.

[30] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, “Spectralformer: Rethinking hyperspectral image classification with transformers,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, Nov. 2021.

[31] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, “Deep convolutional neural networks for hyperspectral image classification,” *J. Sensors*, vol. 2015, pp. 1–12, Jul. 2015.

[32] W. Zhao and S. Du, “Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016.

[33] C. Chen, J.-J. Zhang, C.-H. Zheng, Q. Yan, and L.-N. Xun, “Classification of hyperspectral data using a multi-channel convolutional neural network,” in *Proc. IEEE Int. Conf. Intell. Comput. (ICIC)*, Jul. 2018, pp. 81–92.

[34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] Z. Zhong, J. Li, Z. Luo, and M. Chapman, “Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.

[36] W. Wang, S. Dou, Z. Jiang, and L. Sun, “A fast dense spectral–spatial convolution network framework for hyperspectral images classification,” *Remote Sens.*, vol. 10, no. 7, p. 1068, Jul. 2018.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[38] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, “Deep feature extraction and classification of hyperspectral images based on convolutional neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[39] Y. Li, W. Xie, and H. Li, “Hyperspectral image reconstruction by deep convolutional neural network for classification,” *Pattern Recogn.*, vol. 63, pp. 371–383, Mar. 2017.

[40] 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.

[41] Z. Gong, P. Zhong, Y. Yu, W. Hu, and S. Li, “A cnn with multiscale convolution and diversified metric for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3599–3618, Jun. 2019.

[42] Z. Li, L. Huang, and J. He, “A multiscale deep middle-level feature fusion network for hyperspectral classification,” *Remote Sens.*, vol. 11, no. 6, pp. 695:1–695:20, Mar. 2019.

[43] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double–branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 11, pp. 1307:1–1307:22, Jun. 2019.

[44] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double–branch dual–attention mechanism network," *Remote Sens.*, vol. 12, no. 3, pp. 582:1–582:25, Feb. 2020.

[45] X. Wang, K. Tan, P. Du, C. Pan, and J. Ding, "A unified multiscale learning framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, Feb. 2022.

[46] T. N. Kipf and M. Welling, "Semi–supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2017, pp. 1–14.

[47] F. Yu and V. Koltun, "Multi–scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2016.

[48] S. Wan, C. Gong, P. Zhong, B. Du, L. Zhang, and J. Yang, "Multiscale dynamic graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3162–3177, May 2020.

[49] Y. Ding, X. Zhao, Z. Zhang, W. Cai, and N. Yang, "Multiscale graph sample and aggregate network with context–aware learning for hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens*, vol. 14, pp. 4561–4572, Apr. 2021.

[50] Y. Ding, X. Zhao, Z. Zhang, W. Cai, N. Yang, and Y. Zhan, "Semi-supervised locality preserving dense graph neural network with arma filters and context-aware learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, Aug. 2021.

[51] Y. Ding, Z. Zhang, X. Zhao, W. Cai, N. Yang, H. Hu, X. Huang, Y. Cao, and W. Cai, "Unsupervised self-correlated learning smoothy enhanced locality preserving graph convolution embedding clustering for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Aug. 2022.

[52] Z. Zhang, Y. Ding, X. Zhao, L. Siye, N. Yang, Y. Cai, and Y. Zhan, "Multireceptive field: An adaptive path aggregation graph neural framework for hyperspectral image classification," *Expert Syst. Appl.*, vol. 217, pp. 119 508–119 522, May 2023.

[53] D. Wang, B. Du, and L. Zhang, "Spectral-spatial global graph reasoning for hyperspectral image classification," *IEEE T. Neur. Net. Lear.*, pp. 1–14, May 2023.

[54] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[55] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "Cnn-enhanced graph convolutional network with pixel-and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021.

[56] Y. Ding, Z. Zhang, X. Zhao, D. Hong, W. Cai, C. Yu, N. Yang, and W. Cai, "Multi-feature fusion: Graph neural network and cnn combining for hyperspectral image classification," *Neurocomputing*, vol. 501, pp. 246–257, Aug. 2022.

[57] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 1559–1572, Jan. 2022.

[58] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[59] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "Hsi-bert: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020.

[60] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, Oct. 2021.

[61] C. Zhao, B. Qin, S. Feng, W. Zhu, W. Sun, W. Li, and X. Jia, "Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning," *IEEE Trans. Image Process.*, vol. 32, pp. 3606–3621, Jun. 2023.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, pp. 1–11, Dec. 2017.

[63] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, Sept. 2022.

[64] F. Zhao, J. Zhang, Z. Meng, H. Liu, Z. Chang, and J. Fan, "Multiple vision architectures-based hybrid network for hyperspectral image classification," *Expert Syst. Appl.*, vol. 234, pp. 121 032:1–121 032:16, Dec. 2023.

[65] A. Yang, M. Li, Y. Ding, D. Hong, Y. Lv, and Y. He, "Gtfn: Gcn and transformer fusion network with spatial-spectral features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, Sept. 2023.

[66] Z. Zhao, X. Xu, S. Li, and A. Plaza, "Hyperspectral image classification using groupwise separable convolutional vision transformer network," *IEEE Trans. Geosci. Remote Sens.*, Mar. 2024.

[67] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11 963–11 975.

[68] S. Liu, T. Chen, X. Chen, X. Chen, Q. Xiao, B. Wu, T. Kärkkäinen, M. Pechenizkiy, D. Mocanu, and Z. Wang, "More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity," *arXiv preprint arXiv:2207.03620*, 2022.

[69] X. Ding, Y. Zhang, Y. Ge, S. Zhao, L. Song, X. Yue, and Y. Shan, "Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 5513–5524.

[70] K. W. Lau, L.-M. Po, and Y. A. U. Rehman, "Large separable kernel attention: Rethinking the large kernel attention design in cnn," *Expert Syst. Appl.*, vol. 236, p. 121352, Feb. 2024.

[71] H. Chen, X. Chu, Y. Ren, X. Zhao, and K. Huang, "Pelk: Parameter-efficient large kernel convnets with peripheral convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 5557–5567.

[72] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1–9.

[73] J. T. Barron, "Continuously differentiable exponential linear units," *arXiv preprint arXiv:1704.07483*, 2017.

[74] 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.

[75] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep roots: Improving cnn efficiency with hierarchical filter groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1231–1240.

[76] F. Zhao, J. Zhang, Z. Meng, and H. Liu, "Densely connected pyramidal dilated convolutional network for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 17, p. 3396, Aug. 2021.

[77] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama *et al.*, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[78] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "Whu-hi: Uav-borne hyperspectral with high spatial resolution (h2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with crf," *Remote Sens. Environ.*, vol. 250, p. 112012, Dec. 2020.

[79] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel resnet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sept. 2021.

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

[81] Q. Liu, Y. Dong, Y. Zhang, and H. Luo, "A fast dynamic graph convolutional network and cnn parallel network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, May 2022.

[82] H. Zhou, F. Luo, H. Zhuang, Z. Weng, X. Gong, and Z. Lin, "Attention multi-hop graph and multiscale convolutional fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, Apr. 2023.