Fully Group Convolutional Neural Networks for Robust Spectral-Spatial Feature Learning

Xian Li, *Student Member, IEEE*, Mingli Ding, and Aleksandra Pižurica, *Senior Member, IEEE*

***Abstract*—Convolutional Neural Network (CNN) has been widely applied in hyperspectral image (HSI) classification exhibiting excellent performance. Weak generalization of CNN models to different data sets is a common issue in this domain largely due to limited amount of labelled training samples. In this paper, we propose a *fully* group convolutional neural network (FGCNN) method that integrates cascades of shuffled group convolutions tailored to different network stages. To our knowledge, this is the first reported full group CNN model in general, and we design it in particular for robust spectral-spatial classification of HSI. In the primary feature extraction stage, we develop an original multi-scale spectral feature extraction approach based on a novel concept of multi-kernel depthwise convolution that we define in terms of shuffled and importance-weighted group convolution. In the subsequent stage, we introduce a discriminative spectral-spatial feature extraction method with a novel group competition block to capture informative features with relatively few parameters. The final feature fusion stage, is defined as a novel lightweight group feature fusion method that sharply reduces fusion weights compared to traditional methods with fully connected layers. Experimental results on three data sets show that the proposed FGCNN yields robust classification accuracy under the same hyperparameter settings compared to the current state-of-the-art.**

***Index Terms*—Group convolution, multi-scale spectral feature extraction, spectral-spatial feature learning, lightweight feature fusion, hyperspectral image classification.**

I. Introduction

HYPERSPECTRAL imaging is now established as one of the key technologies in remote sensing [1]. While offering rich spectral information in hundreds of continuous spectral bands, hyperspectral images (HSIs) remain to pose challenges for processing [2, 3] due to their huge dimensionality, and lack of sufficient training data to match it. Feature extraction is a common way to address this challenge.

Recent studies demonstrate the success of deep learning in HSI feature extraction [4–6]. Diverse deep learning models have been applied to spectral feature extraction, including stacked auto-encoders [7], deep belief networks [8], recurrent neural networks [9], one dimensional convolutional neural networks (1D-CNN) [10], and graph convolutional networks [11]. Because of the spatial variability of spectral signatures, spectral noise and other ambiguities in the data, spectral-spatial classification that incorporates spatial context typically outperforms spectral classification alone [12, 13]. Both for extracting spatial features and for combined spectral-spatial features, CNNs are by far the most often used deep learning model [5], typically employed with 2D or 3D neurons [14–16].

Current CNN models often use overparameterized networks to boost the classification performance [17–19]. However, the shortage of labelled pixels leads to a weaker generalization [20, 21]. Data augmentation [22] can mitigate this problem to a certain extent, as well as transfer learning [15], weakly supervised learning [23], few-shot learning [24, 25], generative adversarial learning [26, 27], and neural architecture search [28, 29]. Still, the inherent limitations of the models remain a limiting factor for the performance of the whole system. Recent approaches to reducing the amount of the learning parameters including combining CNNs with traditional machine learning models, such as Gabor filters [30], Markov random fields [31], and conditional random fields [32], which allows them to reduce the number of the convolutional layers. Compared to pure CNNs, these models involve some type of feature engineering [33].

The approach that we follow here is based on group feature extraction [34], which reduces redundancy of the learned weights in deep networks [20]. The term group convolution refers to dividing the input channels into distinct groups and performing a regular convolution over each group separately. A downside is a weaker representation due to ignoring the correlations among the different groups. A clever idea of channel shuffling was recently introduced in computer vision [35], to incorporate correlation across different groups. Related methods include interleaved group convolutions (IGCs) [36–38] and fully learnable group convolution (FLGC) [39]. A limitation of the current group CNN models including ShuffleNet [35], IGCs [36–38], and FLGC [39] methods is that they employ group convolution only partially (i.e., only in some stages of the network), and never for primary feature extraction nor for feature fusion.

In this paper, we develop a *fully* group convolutional neural network (FGCNN). We give a unified and compact mathematical description, where different learning stages share the same operations, tailored to the particular tasks (such as primary and subsequent feature extractions, and feature fusion). This is to the best of our knowledge the first reported fully group CNN model in general, and we design it in particular for robust spectral-spatial classification of HSI. Our architecture

This work was partially supported by the China Scholarship Council, and received funding from the AI program of the Flemish Government.

X. Li is with the School of Instrumentation Science and Engineering, Harbin Institute of Technology, 150001 Harbin, China, and also with the Department of Telecommunications and Information Processing, UGent-GAIM, Ghent University, 9000 Ghent, Belgium (e-mail: xianli0511@gmail.com).

M. Ding is with the School of Instrumentation Science and Engineering, Harbin Institute of Technology, 150001 Harbin, China (e-mail: dingml@hit.edu.cn).

A. Pižurica is with the Department of Telecommunications and Information Processing, UGent-GAIM, Ghent University, 9000 Ghent, Belgium (e-mail: Aleksandra.Pizurica@UGent.be)

is based on 2D-CNN. In contrast to most of the 2D-CNN methods, including [17, 40–42], which need to apply some form of dimensionality reduction (like PCA) prior to feature extraction, our learning architecture consumes the original raw 3D HSI data. We accomplish this through an efficient multi-scale spectral feature extraction method. Its key novelty is a *multi-kernel* depthwise convolution that we define such to weight the spatial information of each band from different scales. We express formally this multi-scale feature extraction operation in terms of shuffled and importance-weighted group convolution operations. The subsequent stage in our model is a discriminative feature extraction approach, which consists of two parallel streams of shuffled group convolutions with different depths and kernel sizes. We introduce element-wise maximum operation to identify automatically informative features of the two streams. The final stage in our model is a novel lightweight feature fusion method, which reduces sharply the number of learning parameters compared to the commonly used fully connected layers while enhancing the feature fusion capability. The whole learning framework is defined in a unified, consistent manner, where the same core principle of shuffled group convolutions is built in the different network stages. The proposed framework proves to be robust to changing characteristics of the datasets and yields not only improved accuracy but also remarkably stable performance.

The main contributions of this work are the following:

1) We propose a *fully* group convolutional network for the classification of hyperspectral images based on spectral-spatial features. To our knowledge, this is the first reported fully group convolutional network in general. The main advantage of the proposed method is its robustness, exhibited as a remarkably stable performance when applied to different data sets.

2) We introduce multi-kernel depthwise convolution to weight the spatial information at different scales in each band. Based on this idea, we develop a multi-scale spectral feature extraction method, which consumes raw hyperspectral data with *all* spectral bands. This is an important asset compared to the current CNN models for HSI classification that often require some form of dimensionality reduction at the input.

3) We propose a novel lightweight feature fusion method that sharply reduces the fusion weights compared to traditional methods with fully connected layers. This lightweight fusion method can be applied to other networks too and is especially of interest when the available training data is rather limited.

4) We devise an effective approach to reduce the number of hyperparameters and the time required for their tuning. This enables automatic adaptation of the proposed classification method to different data sets while maintaining the same hyperparameters settings.

A preliminary version of a part of this work was reported in a conference paper [43]. Here, we build further on this work, and present important novelties and improvements in the performance. Apart from more elaborate theoretical and experimental analysis, and formal presentation, the main differences are the following. Firstly, in contrast to [43], here we build a fully group convolutional neural network (FGCNN), which is also one of the most important contributions of this paper in general. This fully group convolutional design leads to a more elegant formulation and to improved classification performance. Secondly, we introduce here a new lightweight feature fusion method as opposed to the conventional fusion method with fully connected layers that was used in [43]. This novel feature fusion method is important in its own right and can be used independently from the rest of our model, to boost the performance of other CNN classifiers. Thirdly, the discriminative spectral-spatial feature extraction developed here differs significantly from [43], and employs a novel group competition structure with parallel network streams. Furthermore, we give elaborate analysis of our FGCNN model in terms of parameter settings and reducing memory requirements. These novelties led to significantly improved performance compared to [43]: the overall classification accuracy increased by more than 6% on some data sets, and much more stable performance over different data sets is reached.

The rest of this paper is organized as follows. Section II reviews related work on spectral-spatial feature extraction and basic ideas behind group feature learning. Section III presents the proposed method in detail. Section IV evaluates the effectiveness of the proposed approach on real hyperspectral images and Section V draws the conclusion of this work.

II. Related Work

A. Spectral-Spatial Feature Learning

Classification of hyperspectral images based on spectral information alone is susceptible to noise and spatial variability of spectral signatures. Combined spectral-spatial feature learning mitigates these adverse effects by incorporating the spatial information. 3D CNN-based methods [10, 44] extract spectral and spatial information jointly from 3D HSI data. These methods often use a relatively small filter size in the spectral dimension to avoid overfitting. Consequently, the resulting classification maps tend to be oversmoothed [40].

Alternative, multi-stream models [11, 14, 40, 45] extract spectral and spatial features separately, and fuse them subsequently. The spectral stream is then independently designed from the spatial one and can be based on various models, including 1D-CNN [45], recurrent neural networks (RNN) [40], and emerging graph convolutional networks [11]. 1D-CNN models typically extract local spectral features due to local connection mechanism, while RNN models can learn non-adjacent spectral features from sequential perspective. Graph convolutional networks are capable of generalizing the operation of convolution from grid data to graph data by using non-Euclidean structure. Here, we adopt another approach, where spectral-spatial features are extracted via 2D-CNNs. The representative methods of this kind [10, 40–42] typically apply first some form of dimensionality reduction like principal component analysis and use only several first principal components as input. This way, fewer learning parameters are needed and thus the computational cost is reduced. However, the spectral information is less well exploited [10, 40] and the number of principal components is often inconsistent in different data sets [14, 18, 42, 46]. We introduce here a novel multi-scale spectral feature extraction method based on 2D-CNN that requires no dimensionality reduction, making use of all spectral bands.

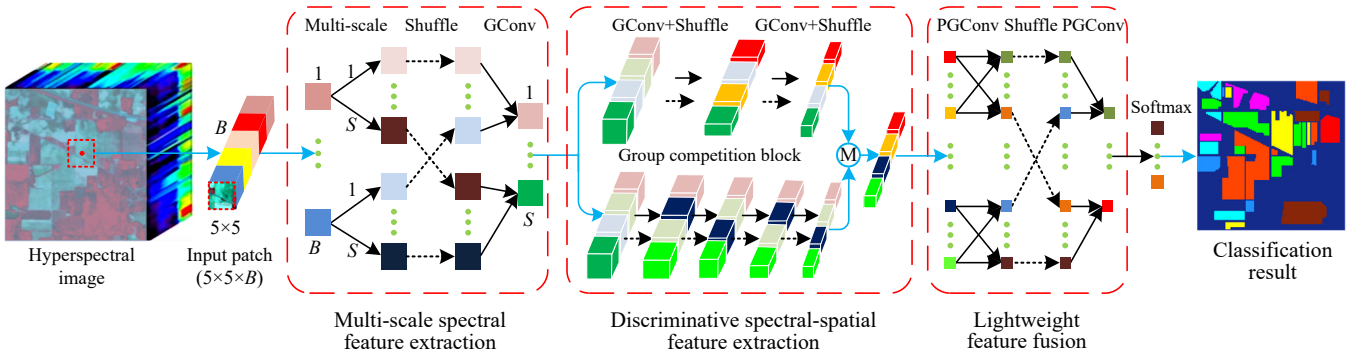

Fig. 1. The overall architecture of the proposed FGCNN. Black solid arrows illustrate here weights and black dashed arrows represent shuffling operations. GConv denotes group convolution operation and PGConv denotes pointwise group convolution operation. Ⓜ represents element-wise maximum operation.

B. Group Feature Learning

Convolutional neural networks yield a highly redundant representation [20]. Group feature learning reduces the redundancy of the representation and the amount of learnable parameters by using group convolution. The term group convolution refers to dividing the input channels into distinct groups and performing a regular convolution operation over each group separately. In HSI processing, group convolution was successfully used both with deep belief networks [47] and with CNN models for band-adaptive feature learning [34]. The methods mentioned above process each group of channels independently, ignoring thus correlation between channels assigned to different groups.

A ShuffleNet method of Zhang *et al*. [35] that was developed specifically for mobile devices, allows for inter-group correlations by channel shuffling. A related approach based on interleaved group convolutions (IGC) was proposed in [36], and later extended to a structured sparse version [37] and a low-rank version [38]. Recently, a FLGC method was proposed to learn the group structure for reducing the computational cost [39]. These methods demonstrated huge success in RGB image processing, but they employ group convolution only in some stages of the network, and never for primary feature extraction nor for feature fusion. Partially, this is because these methods were primarily aimed for RGB images, and a regular convolutional layer is then typically utilized to enlarge the number of channels for the following group convolutions [35–39].

We develop instead a *fully* group convolutional network, which consumes at the input raw 3D data with an arbitrary number of input channels, including RGB images and HSIs with different numbers of spectral bands. Our spectral feature extraction method builds on the concept of depthwise separable convolution [48], which is commonly referred to as "separable convolution" in deep learning frameworks. It consists of a depthwise convolution followed by a pointwise convolution. The depthwise convolution is a spatial convolution performed independently over each channel of an input, saving computation greatly. The pointwise convolution is 1×1 convolution, which projects the output of the depthwise convolution onto a new channel space and enables this way cross-channel correlations. This separable convolution is different from spatially separable convolution. We shall employ both depthwise and spatially separable convolution to introduce a multi-kernel generalization of the depthwise convolution.

For feature fusion, several fully connected layers are typically used both in general computer vision works [35–39] and in hyperspectral image analysis [19, 42, 49]. As opposed to this common fusion approach that involves a large number of fusion parameters, we propose a lightweight fusion method while maintaining the fusion ability.

III. Proposed Method

A. Overall Architecture

A major challenge faced by CNN-based models for HSI classification is a weak generalization to different images. Here, we propose a novel approach to learning spectral-spatial features effectively with a fully group CNN also in the case of relatively few labelled data. This way we achieve much more robust and stable performance, which is evidenced by the results on varying data sets. Earlier reported group CNNs [35–39] employed group convolution only in some stages of the networks and to our knowledge our model that we abbreviate as FGCNN is the first reported fully group CNN model in general. Its first stage extracts multi-scale spectral features from the raw hyperspectral data with all spectral bands, and reduces the spectral dimensionality. The second stage performs a discriminative spectral-spatial feature extraction powered by a novel group competition block. The final stage is a lightweight feature fusion, which contains much fewer fusion parameters compared to traditional methods with fully connected layers. Fig. 1 shows the overall architecture of the proposed FGCNN, with its three main components: 1) multi-scale spectral feature extraction, 2) discriminative spectral-spatial feature extraction, and 3) lightweight feature fusion. In the following, we present the details of the three components and a network design guideline. The definitions of the symbols used in the paper are given in Table I for better readability.

B. Multi-Scale Spectral Feature Extraction

Since HSIs contain hundreds of bands, current 2D-CNN based models often use only few principal components as inputs to reduce the number of the parameters [17, 40–42],

TABLE I
THE SYMBOLS USED IN THIS PAPER AND THEIR DEFINITIONS.

Symbol	Definition	Symbol	Definition
$\mathcal{X}; \mathbf{X}; \mathbf{x}$	HSI patch; matrix; vector	$\mathbf{Y}$	output of the multi-scale spectral feature extraction
$L = P \times P$	spatial size of $\mathcal{X}$	$\mathbf{F}_{l-1}; \mathbf{F}_l$	input and output of the l-th group competition block
B	number of spectral bands of $\mathcal{X}$	$d; l$	number of group competition blocks; block index
$S; s$	number of scales; a particular scale (scale index)	k	number of channels for each competition block
$\mathbf{V}$; $\mathbf{U}$	depthwise convolution; pointwise convolution	MP; $\mathcal{M}$	max pooling; element-wise maximum operator
$\mathbf{V}_S; v_{i,s}$	multi-kernel depthwise convolution matrix; the s-th kernel (weight) for the i-th band of $\mathcal{X}$	$\mathbf{W}^l_{i,j}; \mathbf{P}^l_{i,j}$	weights and shuffling operator of the j-th group convolution for the i-th residual stream in block l
$\mathbf{P}$	shuffling operator (permutation matrix)	$\mathbf{e}^l_i$	importance weights for the i-th stream in block l
$\mathbf{e}; e_{i,s}$	importance weighting matrix; weight along with $v_{i,s}$	$\mathbf{W}^j; \mathbf{P}^j$	weights and shuffling operator for the j-th fusion layer
$\mathbf{U}_S; \mathbf{u}_s$	pointwise group convolution matrix; weights for s-th group	n	number of fusion layers
b	number of adjacent feature maps for learning $e_{i,s}$	c_i	number of channels for fusion layer i
m	number of output channels for each group in $\mathbf{U}_S$	g	number of groups

sacrificing some useful information for the classification [34]. On the other hand, the data-driven feature learning methods that aim to extract more complete spectral features from a higher-dimensional representation face the limitations known as the curse of dimensionality [2].

To mitigate this problem, we propose a novel multi-scale spectral feature extraction approach, based on group convolution, which is able to consume the whole HSI (instead of several principal components). The main idea is to first expand the input channels into their multi-scale representations, with a special kind of group convolution (multi-kernel depthwise convolution) in order to process spectral features at multiple scales simultaneously, and then to feed shuffled and importance-weighted outputs to a group convolution, reducing thereby the output size.

Let $\mathcal{X} \in \mathbb{R}^{P\times P\times B}$ denote an HSI patch with window size of $L = P \times P$ and with B spectral bands, and transform it into a 2D matrix $\mathbf{X} \in \mathbb{R}^{B\times L} = [\mathbf{x}_1^{\mathrm{T}}, \cdots, \mathbf{x}_B^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{x}_i^{\mathrm{T}} \in \mathbb{R}^{1\times L}$ is the i-th band of $\mathcal{X}$. Further on, let us denote separable convolution compactly as

$$\mathbf{Y} = \mathbf{UVX} \tag{1}$$

where the matrix $\mathbf{V}$ acts as operator of the depthwise convolution (i.e., channel-wise spatial convolution) and $\mathbf{U}$ denotes the 1×1 convolution, which maps cross-channel correlations.

We define the proposed multi-scale spectral feature extraction as

$$\mathbf{Y} = \mathbf{U}_S\mathbf{e} \odot \mathbf{PV}_S\mathbf{X} \tag{2}$$

where $\mathbf{e} \odot \mathbf{PV}_S$ is shuffled and importance-weighted multi-kernel depthwise convolution matrix. $\mathbf{V}_S$ is a multi-kernel generalization of depthwise convolution. $\mathbf{P}$ is the permutation matrix that performs channel shuffling, and $\mathbf{e}$ provides importance weighting through the channel-wise product $\odot$. $\mathbf{U}_S$ denotes the 1×1 group convolution. Fig. 2 illustrates this structure and we give a formal description next.

The core component of the proposed spectral feature extraction approach is a novel multi-kernel depthwise convolution operation. What makes it essentially different from the common depthwise convolution [48] is that it has multiple output channels for each input channel, hence the name multi-kernel depthwise convolution. Pointwise multi-kernel depthwise convolution is a special case where the spatial filter size is of 1×1, leading to a scalar weight on each band at each kernel. We employ a pointwise multi-kernel depthwise convolution to weight the spatial information of each band from S scales as shown in the left of Fig. 1. We define 1×1 multi-kernel depthwise convolution matrix as

$$\mathbf{V}_S = \begin{bmatrix} v_{1,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ v_{1,S} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_{B,1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_{B,S} \end{bmatrix} \tag{3}$$

where $[v_{i,1}, \cdots, v_{i,S}]^{\mathrm{T}}$ are the multi-kernel weights with S scales for $\mathbf{x}_i^{\mathrm{T}}$. A common depthwise convolution $\mathbf{V}$ is a special case of the above defined multi-kernel depthwise convolution $\mathbf{V}_S$ for $S = 1$.

The shuffling operator $\mathbf{P}$ shuffles $\mathbf{V}_S\mathbf{X} \in \mathbb{R}^{S\cdot B\times L}$ with B groups:

$$\mathbf{V}_S\mathbf{X} = [\underbrace{v_{1,1}\mathbf{x}_1^{\mathrm{T}}, \cdots, v_{1,S}\mathbf{x}_1^{\mathrm{T}}}_{\text{1st band}}, \cdots, \underbrace{v_{B,1}\mathbf{x}_B^{\mathrm{T}}, \cdots, v_{B,S}\mathbf{x}_B^{\mathrm{T}}}_{B\text{-th band}}]^{\mathrm{T}} \tag{4}$$

into $\mathbf{PV}_S\mathbf{X} \in \mathbb{R}^{B\cdot S\times L}$ with S groups:

$$\mathbf{PV}_S\mathbf{X} = [\underbrace{v_{1,1}\mathbf{x}_1^{\mathrm{T}}, \cdots, v_{B,1}\mathbf{x}_B^{\mathrm{T}}}_{\text{1st shuffled group}}, \cdots, \underbrace{v_{1,S}\mathbf{x}_1^{\mathrm{T}}, \cdots, v_{B,S}\mathbf{x}_B^{\mathrm{T}}}_{S\text{-th shuffled group}}]^{\mathrm{T}} \tag{5}$$

where $[v_{1,s}\mathbf{x}_1^{\mathrm{T}}, \cdots, v_{B,s}\mathbf{x}_B^{\mathrm{T}}]^{\mathrm{T}}$ is the s-th shuffled group, which contains all the spectral bands at a particular scale s ($1 \le s \le S$).

Not all the feature maps are equally informative, and it is thus of interest to identify the most important ones and to suppress the others by some proper weighting. We accomplish this by an operation that generalizes the so-called squeeze-and-excitation (SE) operation [44], which learns automatically the appropriate channel importance weights $\mathbf{e} \in \mathbb{R}^{B\cdot S\times 1} = [e_{1,1}, \cdots, e_{B,1}, \cdots, e_{1,S}, \cdots, e_{B,S}]^{\mathrm{T}}$ from their feature maps $\mathbf{PV}_S\mathbf{X} \in \mathbb{R}^{B\cdot S\times L}$. Since the adjacent bands have strong

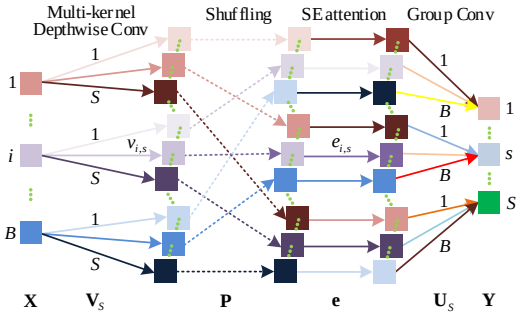

Fig. 2. An illustration of the proposed multi-scale spectral feature extraction.

correlations, we extend this SE operation to a *group* SE that lets the adjacent feature maps share the same channel importance weight: $e_{i,s} = \cdots = e_{i+b,s}$, where b is the number of adjacent feature maps. This way, each channel weight $e_{i,s}$ in our group SE is learned automatically from the corresponding group of feature maps: $[v_{i,s}\mathbf{x}_i^{\mathrm{T}}, \cdots, v_{i+b,s}\mathbf{x}_{i+b}^{\mathrm{T}}]$. The amount of the parameters for our group SE drops by a factor b compared to a regular SE.

Each shuffled and importance-weighted group is then fed to the 1×1 convolution separately, to extract global spectral features. We define the group convolution matrix as

$$\mathbf{U}_S = \begin{bmatrix} \mathbf{u}_{1,1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{u}_{S,S} \end{bmatrix} \tag{6}$$

where $\mathbf{u}_{s,s} \in \mathbb{R}^{m \times B}$ denotes the weights of the s-th group. m is the number of output channels for each group, which is set to 1 in our experiment. Each group extracts global (all) spectral features $\mathbf{u}_{s,s}[e_{1,s}v_{1,s}\mathbf{x}_1^{\mathrm{T}}, \cdots, e_{B,s}v_{B,s}\mathbf{x}_B^{\mathrm{T}}]$ at a particular scale s. The net result amounts to extracting S multi-scale spectral features. The expression (2) can be written as

$$\mathbf{Y} = \mathbf{U}_S\mathbf{e} \odot \mathbf{P}\mathbf{V}_S\mathbf{X} = \mathbf{W}\mathbf{X} \tag{7}$$

where $\mathbf{W} = \mathbf{U}_S\mathbf{e} \odot \mathbf{P}\mathbf{V}_S$ is a composite convolution kernel consisting of the shuffled and importance-weighted multi-kernel depthwise convolution operator $\mathbf{V}_S$ and the group convolution operator $\mathbf{U}_S$. Fig. 2 illustrates these operators. The proposed multi-scale spectral feature extraction method behaves as a kind of inverted bottleneck [50] where the number of channels is firstly expanded (from B to BS) and then squeezed (from BS to mS, $m \ll B$). A large S consumes more memory due to a huge number of channels (i.e., BS). We reduce the memory requirements by extending our spectral feature extraction method to a group version that divides S into several groups and performing our spectral feature extraction method over each group separately.

C. Discriminative Spectral-Spatial Feature Extraction

Making use of the extracted multi-scale spectral features, we want to build next more discriminative spectral-spatial features. We also want to keep a unified, consistent framework that shares the same building blocks across the whole learning system. To this end, we extend further the core component from equation (7), employing it now within a spectral-spatial framework.

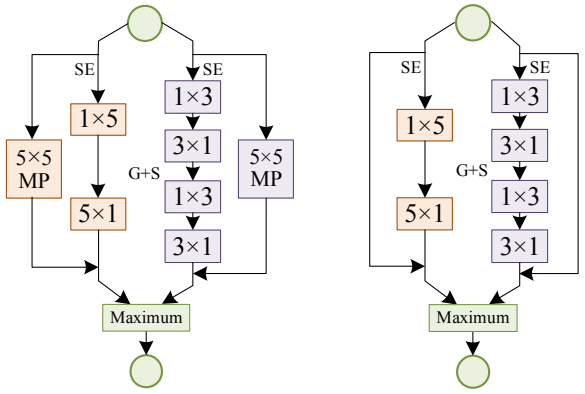

Fig. 3. The architectures of (a) group competition block with no padding, (b) group competition block with padding. G+S illustrates shuffled group convolution operation. We employ spatial separable convolution that factorizes $r * r$ (r=3 or 5) into $1 \times r$ and $r \times 1$. MP represents a max pooling operation. Maximum represents the element-wise maximum operation. SE represents the squeeze-and-excitation operation.

The core component of the proposed approach is a novel multi-stream module that we call group competition block. Fig. 3 shows its two versions, for larger and smaller input sizes, the details of which are explained later. The proposed group competition block consists of two parallel residual streams of shuffled group convolutions, which have the same receptive field but different depths and kernel sizes. While current multi-stream blocks including Inception [51], typically employ concatenation operation and 1×1 convolution operation to fuse multi-stream features, we introduce element-wise maximum operation, to automatically identify informative features of the two streams. An advantage of this approach is that we avoid increasing the output dimension and the additional parameters. The two streams based on group convolution perform element-wise feature selection automatically. In each stream, we exploit residual learning to facilitate the training process. In contrast to ShuffleNet [35] and IGC [36], which use extensively 1×1 convolutions which are known to incur some loss in accuracy [52], we avoid 1×1 convolutions. Instead, we employ spatially separable convolution that factorizes convolutions of filter size $r \times r$ to a combination of $1 \times r$ and $r \times 1$ and then we apply shuffled group convolution to these. This way, the number of parameters is reduced. Shuffling enables inter-group correlations and importance weighting promotes most informative features to be learned.

To adapt to different input spatial size, we design group competition block both with no padding for a large input spatial size as shown in Fig. 3 (a) and with padding for a small one as shown in Fig. 3 (b).

Let $\mathbf{F}^{l-1}$ be the input of the l-th group competition block, $1 \leq l \leq d$. The input to the first block is the output of the previous network stage: $\mathbf{F}^0 = \mathbf{Y}$. We define the output of i-th residual stream as

$$\mathbf{F}_i^l = \mathbf{C}_i^l\mathbf{e}_i^l \odot \mathbf{F}^{l-1} + \mathrm{MP}(\mathbf{F}^{l-1}) \tag{8}$$

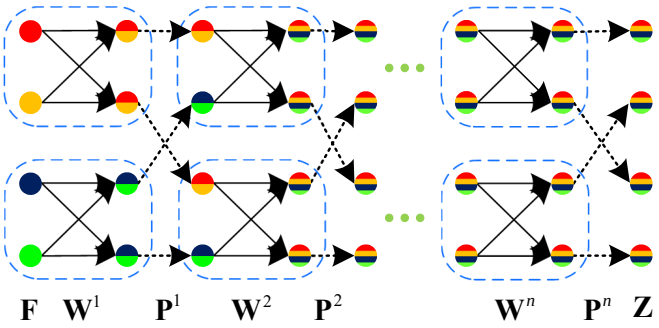

Fig. 4. An illustrated architecture of the proposed lightweight feature fusion. Dashed boxes denote pointwise group convolutions, solid arrows denote weights, dashed arrows denote shuffling operations. The colors of circles show the fusion process of the input $\mathbf{F}$. Each fused feature is related to all the input features in $\mathbf{F}$ after $\mathbf{W}^2$.

where

$$\mathbf{C}_i^l = \prod_j \mathbf{P}_{i,j}^l \mathbf{W}_{i,j}^l \tag{9}$$

is the cascade of shuffled group convolution for the i-th stream in the l-th group competition block. $\mathbf{e}_i^l$ denotes importance weighting operation, and MP is the max pooling operation. $\mathbf{W}_{i,j}^l$ and $\mathbf{P}_{i,j}^l$, $i = \{1, 2\}$, $j = 1, \cdots, 2i$ are the weights and shuffling operator of the j-th group convolution for the i-th stream in the l-th group competition block. By introducing the element-wise maximum operation $\mathcal{M}$, we define the output of the l-th group competition block as

$$\mathbf{F}^l = \mathcal{M}(\mathbf{F}_1^l, \mathbf{F}_2^l) \tag{10}$$

The output for $l = d$, is the output of the proposed discriminative spectral-spatial feature extraction stage $\mathbf{F} = \mathbf{F}^d$.

D. Lightweight Feature Fusion and Classification

We extend now the above introduced approach based on shuffled pointwise group convolutions also to feature fusion in the final stage of the learning architecture. This leads to a lightweight feature fusion method that reduces sharply the number of fusion parameters compared to the existing methods that use several fully connected layers.

In particular, we fuse the input features through a cascade of shuffled pointwise group convolutions, as illustrated in Fig. 4. Due to a structured sparse connection, much less parameters are used than with fully connected layers. The amount of the parameters drops by a factor g, where g is the number of groups. By introducing channel shuffling in between the pointwise group convolutional layers, we enable also inter-group cross-correlations and thus more general fusion features, while keeping a relatively small total amount of the fusion parameters. This way, the proposed lightweight feature fusion requires fewer fusion parameters while enhancing the feature fusion capability compared to the fully connected layers with the same units. To our knowledge, feature fusion based on group convolution has not been explored before.

Formally, given the input $\mathbf{F}$, we define the proposed lightweight feature fusion as

$$\mathbf{Z} = (\mathbf{P}^n \mathbf{W}^n) \cdots (\mathbf{P}^j \mathbf{W}^j) \cdots (\mathbf{P}^1 \mathbf{W}^1) \mathbf{F} \tag{11}$$

TABLE II
THE HYPERPARAMETERS OF THE PROPOSED NETWORK FOR ALL THE DATA SETS. g IS THE NUMBER OF GROUPS FOR ALL THE FEATURE LEARNING STAGE. S IS THE NUMBER OF SCALES IN THE PRIMARY STAGE. k AND d REFER TO THE NUMBER OF CHANNELS AND LAYERS IN THE MIDDLE STAGE, RESPECTIVELY. n, c_1 AND c_2 DENOTE THE NUMBER OF FUSION LAYERS AND CHANNELS IN EACH LAYER, RESPECTIVELY. WE ONLY NEED TO TUNE THREE HYPERPARAMETERS: g, d AND n.

Parameter	g	S	k	d	n	c_1	c_2
Value	6	6^2	6^2	1	2	$2{\cdot}6^2$	$1{\cdot}6^2$

here $\mathbf{W}^j$, $1 \leq j \leq n$, is a structured sparse matrix (see equation (6)) representing the weights for the j-th fusion layer, $\mathbf{P}^j$ is the shuffling operator, and n is the number of fusion layers. When $n \geq 2$, each fused feature is related to all the input features in $\mathbf{F}$ (see the colors of circles in Fig. 4) to fuse more general features. To avoid overfitting, we use a dropout with 0.3 threshold before feature fusion and a L2 regularization with 0.2 in the softmax layer. We employ the mini-batch Adadelta [53] to optimize the cross-entropy loss function.

E. Fully Group CNN Design

Having developed each feature learning stage of the proposed FGCNN, we need to determine the hyperparameters. Our goal is to design a robust architecture that can be applied to any data set, and to reduce the number of hyperparameters and the time required for their tuning. To this end, we design a guideline that simplifies the process of hyperparameter setting and tuning on the premise for each data set.

The main idea is to relate the network hyperparameters to the number of groups g. Regarding the number of channels, it is natural to start from the design criteria for the middle spectral-spatial extraction stage that connects the front (primary feature extraction) and the end (fusion) stage. To simplify the optimization process, we fix the number of channels for all the convolutional layers in this middle stage to $k = g^2$, which guarantees that the input channels of each layer can be uniformly divided. Further on, reasoning that the number of scales S in the multi-scale spectral feature extraction stage should be equal to the number of output channels of this stage (which are inputs to the middle stage), we have that $S = g^2$.

In the fusion stage, we let the number of neurons gradually decrease from one layer to the next, as it is commonly done. In particular, we set the number of channels for j-th fusion layer to $c_j = \alpha(n - j)g^2$, where n is the number of fusion layers and $\alpha \geq 1$ is a growth factor, which depends on the amount of training data. In cases where the amount of training data is rather limited (the case that we are interested in), $\alpha = 1$ provides the best performance based on our empirical study. A larger α is recommended in computer vision fields where more training data is available. In summary, we only need to tune three hyperparameters: the number of groups g, the number of group competition blocks d and the number of fusion layers n. We use the same hyperparameters for our network on each data set as shown in Table II. An ablation study regarding the robustness of the hyperparameters is given in Section IV-C.

TABLE III
COMPARISON OF THE CLASSIFICATION ACCURACIES AMONG THE PROPOSED FGCNN AND THE BASELINES USING INDIAN PINES IMAGE.

ID	Train/Test	MRFCNN	PPFCNN	DRCNN	SSRN	HybridSN	ADGAN	MCNNCP	FRCNN	GCNN	**FGCNN**
1	50/1378	69.75±5.61	80.70±3.40	77.56±4.69	83.17±6.75	79.48±5.52	85.65±5.02	81.21±4.54	90.71±3.90	87.16±5.24	**94.04±2.12**
2	50/780	79.22±4.72	79.54±6.60	86.20±5.47	75.85±10.52	87.14±4.13	77.81±12.03	89.88±4.50	94.69±4.80	86.03±4.77	**95.45±4.00**
3	50/433	95.20±1.55	92.44±3.02	96.92±1.05	96.50±2.79	95.84±2.27	90.46±9.21	96.19±1.74	98.01±1.71	95.78±1.69	**98.11±1.59**
4	50/428	99.84±0.11	99.98±0.07	99.77±0.34	99.03±1.50	**100±0**	99.49±0.61	**100±0**	**100±0**	**100±0**	**100±0**
5	50/922	80.38±3.06	75.77±5.17	85.75±3.99	70.20±14.71	86.69±6.26	77.10±9.56	88.81±4.48	90.09±6.5	86.31±7.22	**94.47±3.61**
6	50/2405	65.42±3.36	90.06±2.56	69.18±3.16	87.46±3.78	69.22±6.03	**91.97±2.98**	73.45±8.57	85.85±5.25	79.85±7.87	89.33±3.34
7	50/543	74.00±4.46	81.54±6.65	88.57±4.20	64.86±15.79	83.90±4.88	**99.23±1.10**	92.32±6.10	94.33±3.45	94.77±2.40	97.79±1.13
8	50/1215	96.11±1.61	99.52±0.20	99.04±0.77	99.01±0.52	97.89±1.91	99.30±0.70	99.57±0.46	99.65±0.48	99.65±0.42	**99.82±0.14**
AA	-	82.49±1.30	87.44±1.40	87.87±0.84	84.51±2.28	87.52±1.32	90.13±2.68	90.18±0.81	94.16±1.23	91.19±0.80	**96.12±0.89**
OA	-	77.77±1.72	86.62±1.36	83.23±1.23	81.94±3.15	83.00±1.37	89.74±2.18	85.90±2.11	92.03±1.39	88.31±1.28	**94.48±1.18**
κ	-	73.67±1.97	84.00±1.61	80.15±1.42	78.59±3.61	79.79±1.59	87.58±2.67	83.22±2.39	90.43±1.66	85.98±1.46	**93.37±1.40**

TABLE IV
COMPARISON OF THE CLASSIFICATION ACCURACIES AMONG THE PROPOSED FGCNN AND THE BASELINES USING THE PAVIAU IMAGE

ID	Train/Test	MRFCNN	PPFCNN	DRCNN	SSRN	HybridSN	ADGAN	MCNNCP	FRCNN	GCNN	**FGCNN**
1	50/6581	89.60±0.72	97.41±1.58	93.06±2.33	**99.27±0.61**	93.30±1.57	72.82±5.87	87.38±4.73	82.68±21.14	93.02±2.19	95.74±2.55
2	50/18599	89.83±1.50	97.00±1.30	95.20±2.49	**99.35±0.32**	94.73±2.80	81.01±5.78	83.66±4.20	92.17±4.52	91.13±4.54	95.70±2.48
3	50/2049	86.99±0.54	83.53±4.22	**91.28±2.15**	76.40±13.24	87.99±10.00	90.56±5.08	84.13±6.62	79.63±9.80	80.46±3.28	85.00±3.45
4	50/3014	95.63±0.94	80.34±7.58	96.10±1.22	91.65±11.11	94.26±2.76	93.03±2.95	93.13±2.14	94.40±2.30	95.78±1.96	**96.26±1.28**
5	50/1295	99.61±0.63	99.76±0.44	99.84±0.22	99.95±0.17	**100±0.0**	99.78±0.41	**100±0**	99.99±0.03	99.99±0.04	**100±0.0**
6	50/4979	82.95±1.71	73.65±4.48	93.26±2.89	88.04±4.72	94.21±5.26	87.38±4.21	88.67±7.14	96.55±2.69	94.93±4.31	**96.64±2.64**
7	50/1280	91.70±0.77	86.25±7.61	96.77±1.43	90.00±6.16	98.38±0.44	95.97±2.39	**98.51±0.46**	97.77±1.11	97.34±1.37	98.21±1.07
8	50/3632	80.31±2.53	86.51±6.21	**94.12±2.35**	87.75±4.16	90.41±5.82	73.73±13.14	86.63±5.26	81.17±20.42	85.69±6.32	89.93±2.89
9	50/897	98.72±0.87	96.98±2.93	99.82±0.19	99.87±0.12	99.59±0.47	98.71±1.20	99.22±0.50	99.79±0.25	99.82±0.20	**99.92±0.08**
AA	-	90.59±0.39	89.05±1.67	**95.50±0.46**	92.47±2.60	94.73±1.11	88.11±2.25	91.26±1.21	91.57±2.61	93.13±0.82	95.27±0.42
OA	-	88.99±0.62	90.12±1.56	94.72±0.93	94.19±1.96	94.04±1.13	82.58±3.07	87.06±1.91	90.39±4.05	91.86±2.05	**95.14±1.11**
κ	-	85.56±0.77	87.14±1.99	93.05±1.18	92.38±2.52	92.14±1.45	77.84±3.70	83.27±2.38	87.46±5.12	89.36±2.59	**93.58±1.44**

IV. EXPERIMENTAL RESULTS AND ANALYSIS

We perform experiments on three well-known HSI data sets: Indian Pines, the University of Pavia (denoted as PaviaU) and Salinas. Three objective metrics, overall accuracy (OA), average accuracy (AA), and Kappa coefficient (κ) are used for evaluation. For each experiment, we report the mean and standard deviation of the classification results over ten runs with randomly selected training samples.

A. Data Set Description and Parameter Setting

The Indian Pines image, captured by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor over the agricultural Indian Pines site in northwestern Indiana in 1992, contains 145×145 pixels with 224 spectral bands covering the spectral range from 0.4 to 2.5 μm with a spatial resolution 20 m. It contains 16 ground-truth classes, out of which we select 8 large classes [22, 49], 4 water absorption bands were removed. The PaviaU image, acquired by the ROSIS-03 sensor over an urban area surrounding the University of Pavia, Pavia, Italy, consists of 610×340 pixels with 9 classes and 103 spectral bands covering the spectral range from 0.43 to 0.86 μm with a spatial resolution of 1.3 m. The Salinas image, collected by the AVIRIS sensor over the area of Salinas Valley, CA, USA, has 512×217 pixels with 224 spectral bands covering the spectral range from 0.4 to 2.5 μm with spatial resolution of 3.7 m, 20 water absorption bands were removed.

We randomly select 50 labelled samples per class for training the proposed method from scratch. The remaining labelled samples are used as the test set to evaluate the classification performance. In order to treat boundary pixels in the same way as others in the classification procedure, we apply first mirror-padding, i.e., we extend the HSI with mirror reflections along it boundaries. We then slide a fixed-size window along the padded HSI, extracting the image patches in the same way for all the pixels. We randomly select 10% of the training samples as the validation set to determine the hyperparameters. The hyperparameters of the proposed method are set the same for all the data sets as follows. The initial learning rate is empirically set to 9. The number of training epochs and batch size are empirically set to 300 and 64, respectively. The network hyperparameters of the proposed method are given in Table II. The proposed network is implemented in Keras[1] and TensorFlow[2] framework with Python language.

B. Baseline Comparison

We compare the performance of the proposed FGCNN with the following state-of-the-art CNN-based methods: CNN combined with MRF (MRFCNN) [31], CNN with pixel-pair features (PPFCNN) [22], diverse region based on CNN (DRCNN) [49], spectral-spatial residual network based on 3D-CNN (SSRN) [44], spectral-spatial 3D-CNN followed by spatial 2D-CNN (HybridSN) [54], generative adversarial network with adaptive dropBlock (ADGAN) [17], mixed CNN with covariance pooling for spectral-spatial classification (MCNNCP) [42]. We use spatial patch size 5×5 for all the methods except for ADGAN, which requires patch size 27×27 as explained in [17]. All other parameters of the reference methods are set to

[1]https://keras.io/
[2]https://www.tensorflow.org/

TABLE V
COMPARISON OF THE CLASSIFICATION ACCURACIES AMONG THE PROPOSED FGCNN AND THE BASELINES USING THE SALINAS IMAGE

ID	Train/Test	MRFCNN	PPFCNN	DRCNN	SSRN	HybridSN	ADGAN	MCNNCP	FRCNN	GCNN	**FGCNN**
1	50/1959	99.26±1.33	99.88±0.29	98.15±3.23	99.98±0.04	99.99±0.02	**100±0**	99.75±0.42	99.59±0.38	99.84±0.22	99.96±0.11
2	50/3676	97.65±1.12	99.44±0.31	99.11±1.24	99.75±0.49	99.77±0.27	76.93±29.30	99.84±0.22	99.50±0.65	99.97±0.08	**99.98±0.06**
3	50/1926	98.90±1.38	95.49±3.42	99.31±0.94	98.49±1.31	99.69±0.27	98.12±3.13	99.57±0.86	99.22±0.81	99.97±0.08	**100±0**
4	50/1344	99.50±0.44	96.61±1.27	99.89±0.05	98.88±0.77	99.29±0.45	**99.90±0.20**	99.55±0.79	99.11±0.61	99.82±0.13	99.29±0.45
5	50/2628	96.48±6.80	99.03±1.10	92.31±4.77	**99.36±0.39**	98.96±0.60	98.46±1.42	98.04±1.29	97.46±2.11	98.39±0.45	98.40±0.67
6	50/3909	99.49±0.62	99.86±0.07	99.91±0.08	**99.99±0.02**	99.84±0.22	**99.99±0.01**	99.93±0.11	99.50±0.54	99.94±0.09	99.92±0.12
7	50/3529	98.98±0.81	99.32±1.41	99.16±0.38	**99.99±0.02**	99.85±0.36	95.38±6.46	99.86±0.33	99.84±0.17	99.96±0.04	99.98±0.02
8	50/11221	74.69±4.95	84.82±2.06	82.76±5.68	**87.51±4.43**	78.84±8.18	63.29±23.25	80.14±6.84	81.33±18.05	82.45±5.34	86.27±4.38
9	50/6153	97.48±2.41	99.15±0.35	99.21±0.29	99.57±0.23	99.98±0.05	98.68±1.89	99.63±0.26	99.77±0.20	99.97±0.04	**99.99±0.01**
10	50/3228	92.17±2.46	87.71±3.45	93.69±1.52	97.03±1.46	96.21±1.64	94.99±4.99	95.79±1.75	96.73±1.75	98.15±1.12	**98.37±1.35**
11	50/1018	98.36±1.20	88.55±7.96	99.05±0.24	96.90±2.78	99.64±0.64	96.03±9.27	99.75±0.20	98.13±1.61	**99.97±0.05**	99.95±0.12
12	50/1877	99.90±0.25	98.60±1.02	**100±0**	98.83±1.04	98.96±3.13	95.31±8.04	99.45±0.62	99.16±1.00	99.99±0.02	99.99±0.04
13	50/866	99.58±0.77	98.36±1.67	**100±0**	99.71±0.40	99.76±0.29	98.53±2.28	99.04±0.96	99.75±0.31	99.50±0.40	99.64±0.35
14	50/1020	96.68±1.47	90.98±7.67	97.23±0.68	98.73±0.96	99.16±0.76	**99.39±0.72**	98.80±0.94	98.30±1.94	99.14±0.95	99.24±0.80
15	50/7218	75.82±4.97	71.99±6.15	73.42±12.42	69.52±4.91	86.04±7.09	**94.04±3.31**	83.35±8.11	57.87±29.13	84.29±4.68	85.88±4.18
16	50/1757	97.31±1.52	98.55±0.95	98.92±0.18	**99.68±0.29**	99.08±0.56	96.82±3.67	99.17±0.27	99.10±0.57	99.28±0.34	99.37±0.36
AA	-	95.14±0.65	94.27±1.19	95.76±0.60	96.50±0.22	97.19±0.29	94.12±2.86	96.98±0.60	95.27±1.22	97.54±0.33	**97.92±0.24**
OA	-	89.92±0.87	90.97±1.42	91.66±0.72	92.03±0.66	93.21±0.99	88.59±4.88	93.02±1.32	89.74±2.62	93.92±0.96	**94.96±0.75**
κ	-	88.80±0.96	89.96±1.57	90.72±0.82	91.14±0.72	92.46±1.09	87.40±5.31	92.24±1.46	88.55±2.93	93.23±1.07	**94.39±0.83**

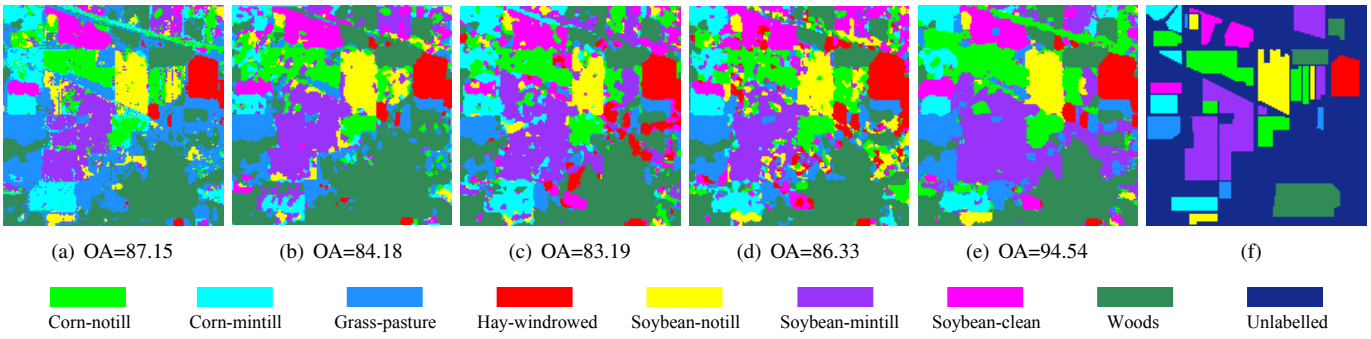

Fig. 5. Classification maps on the Indian Pines image obtained by, (a) PPFCNN, (b) SSRN, (c) HybridSN, (d) MCNNCP, (e) FGCNN, and (f) Ground Truth.

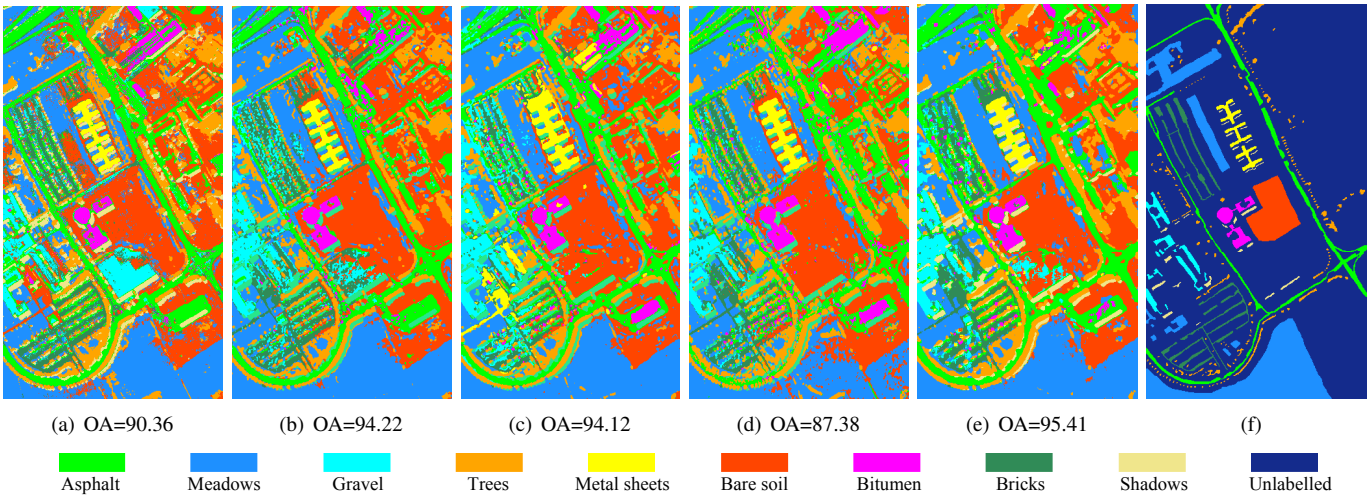

Fig. 6. Classification maps on the PaviaU image obtained by, (a) PPFCNN, (b) SSRN, (c) HybridSN, (d) MCNNCP, (e) FGCNN, and (f) Ground Truth.

the values indicated in the original works. We also compare the proposed FGCNN to its two reduced versions: FGCNN using regular convolution with the same network width (FRCNN) and our conference version (GCNN) [43].

Tables III-V report the classification results of the tested methods on the three data sets. The proposed FGCNN consistently yields the best OA and κ over the reference methods for all the three data sets. For example, on Indian Pines (Table III), the improvement in OA compared to MRFCNN, PPFCNN, DRCNN, SSRN, HybridSN, ADGAN, and MCNNCP methods is about 16.7%, 7.8%, 11.2%, 12.5, 11.5%, 4.7%, and 8.6%, respectively. The gains in OA compared to the best baseline

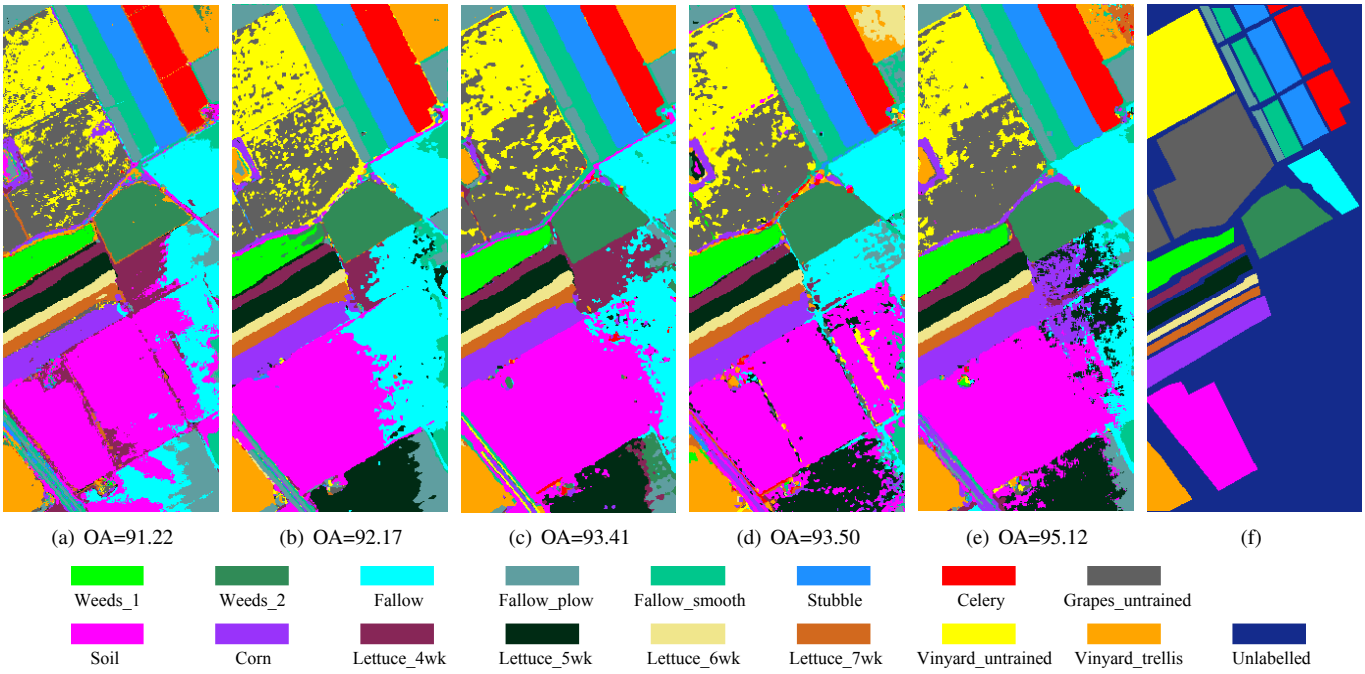

Fig. 7. Classification maps on the Salinas image obtained by, (a) PPFCNN, (b) SSRN, (c) HybridSN, (d) MCNNCP, (e) FGCNN, and (f) Ground Truth.

method are approximately 4.7%, 0.4%, and 1.7% for the Indian Pines, PaviaU, and Salinas images, respectively. It is also evident that our two reduced versions (FRCNN and GCNN) perform better or yield comparable results as the reference methods on the three data sets, which verifies the effectiveness of the proposed architecture. The proposed FGCNN performs consistently better than FRCNN due to the benefit of the proposed full group convolution network. Also, the proposed FGCNN performs better than our earlier conference version GCNN due to the new discriminative spectral-spatial feature extraction and more efficient lightweight fusion. On one test image, PaviaU, DRCNN yields slightly higher AA than the proposed method, but our method yields better OA and κ in this case as well.

In terms of the class-specific accuracy, our FGCNN performs best or yields comparable results to the best ones in most of the classes for the three data sets. Only in several classes this is not the case. For instance, in the PaviaU image, some 'Gravel' (ID=3) samples are misclassified as 'Bricks' (ID=8) due to their huge spectral similarity and the large within-class variation in their spectral reflectance. Figs. 5-7 show the classification maps obtained by different methods on the three data sets. Visually, they are consistent with the results reported in Tables III-V. Obviously, the proposed FGCNN exhibits less noisy estimations compared to reference methods.

We also compare the proposed FGCNN with three methods designed for small-scale training data: DSFL [55], AML [46], and NLGCN [13], as well as two pixel-based methods: CRNN [9] and miniGCN [11] (our FGCNN inputs pixel vectors by discarding the middle spectral-spatial feature extraction in this case). The comparison results are given in Table VI, where the results of the comparative methods are taken from the original works (except for CRNN), and for the proposed FGCNN we show in brackets the results obtained with the same training data as in the corresponding comparative methods. As can be observed in Table VI, our FGCNN consistently yields better OA than the comparative methods for all the three data sets.

TABLE VI
OA OBTAINED BY SEVERAL METHODS DESIGNED FOR SMALL TRAINING DATA AND FOR SPECTRAL CLASSIFICATION. THE RESULTS OF THE PROPOSED FGCNN ARE IN BRACKETS.

Image	Method	Training set	OA
Indian Pines	DSFL	200 samples per class	98.35% (**99.24%**)
	AML	5% per class	79.11% (**94.83%**)
	NLGCN	695 samples in total	87.92% (**94.14%**)
	CRNN	50 pixels per class	69.57% (**72.80%**)
	miniGCN	695 pixels in total	71.97% (**76.23%**)
PaviaU	DSFL	200 samples per class	98.62% (**99.44%**)
	AML	1% per class	89.81% (**94.57%**)
	NLGCN	3921 samples in total	90.04% (**99.54%**)
	CRNN	50 pixels per class	72.30% (**76.54%**)
	miniGCN	3921 pixels in total	77.99% (**92.55%**)
Salinas	DSFL	200 samples per class	98.81% (**99.54%**)
	AML	1% per class	91.63% (**95.46%**)
	NLGCN	50 samples per class	92.48% (**94.96%**)
	CRNN	50 pixels per class	84.61% (**86.43%**)

C. Robustness Analysis and Parameter Tuning

1) Performance on different data sets: By comparing the classification results for the three different data sets in Tables III-V, we observe that the proposed FGCNN shows much more stable performance than the reference methods. While for the five best performing reference methods: DRCNN, SSRN, HybridSN, ADGAN and MCNNCP, OA differs up to 7% from one set to another, for our method this variation is only 1%. This indicates better robustness of the proposed FGCNN to

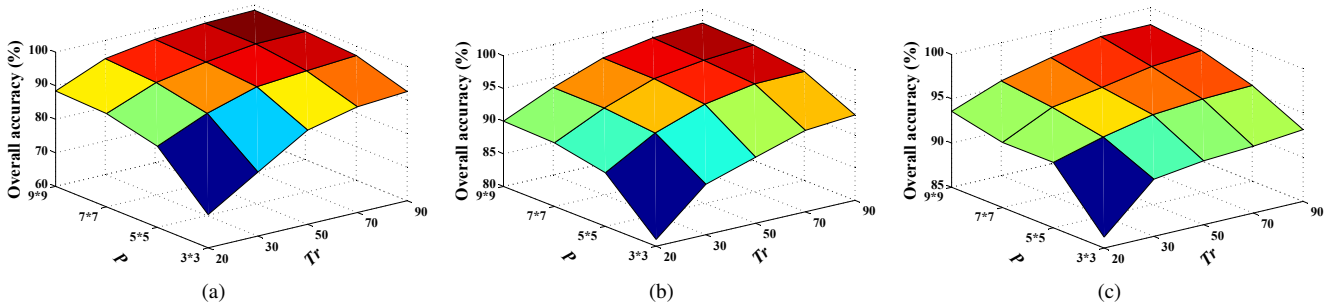

Fig. 8. The overall accuracy of the proposed method with different sizes of patches P and with different numbers of training samples per class Tr (a) Indian Pines data set, (b) PaviaU data set, (c) Salinas data set.

different data sets under the same network settings (Table II), which is an important asset for its practical applicability.

2) Robustness to different volume of training data: We also analyze the influence of the amount of training samples per class Tr and input patch size P on the performance of the proposed method. Fig. 8 shows the mean OA values versus Tr and P for each of the three data sets. As expected, the classification accuracy improves when Tr and P increase. The performance is robust with respect to Tr for all the data sets, especially when $P \geq 5$. Apparently, a larger P which provides richer spatial information yields better mean OA. This comes at a price of increased computational cost and memory requirements.

3) Effect of the number of groups: The number of groups g is a key hyperparameter in our method, which determines the number of channels for all the layers (see Section III-E). Apparently, a larger g increases the memory requirements due to multi-kernel depthwise convolution. We in parallel split our primary spectral feature extraction method into three groups to reduce the memory requirements when $g \geq 7$. The results in Fig. 9 show that the overall accuracy first drastically increases ($g \leq 6$) and then slightly declines or tends to be stable ($g \geq 6$) when g increases, which holds a similar trend for the OA of a regular convolutional layer when the number of channel increases. The main reason is that a smaller g underfits the features and an excessive S tends to overfit them. On PaviaU image, the best overall accuracy has a larger g compared to the other two images due to a smaller number of spectral bands B (103, 220, and 204 for the PaviaU, Indian Pines, and Salinas images, respectively). Since a smaller B requires less model parameters at the same g. We choose $g = 6$ as a tradeoff between the accuracy and the memory requirements for the three data sets.

4) Analysis of the multi-scale spectral feature extraction stage: It is also of interest to analyze the influence of the number of scales S within our multi-scale spectral feature extraction method on the overall performance. By setting the other parameters to the values from Table II, we perform the classification experiments with different S. The results in Fig. 10 show that the overall accuracy first increases with increasing S and then tends to be stable or slightly declines for the PaviaU and Salinas images. For the PaviaU and Salinas data sets the performance remains stable over a broad range of S values. For the Indian Pines image, the overall accuracy

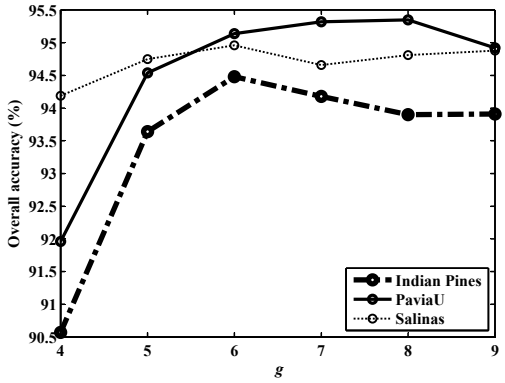

Fig. 9. The influence of the number of groups g on the overall classification accuracy in three data sets.

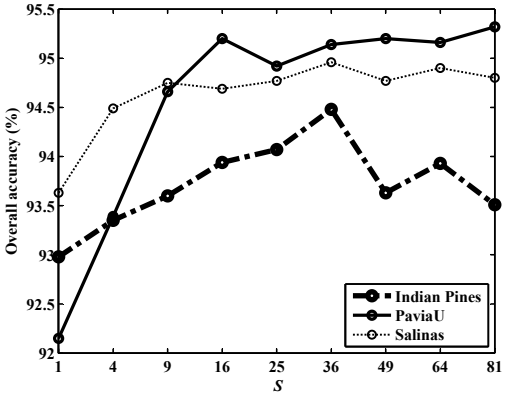

Fig. 10. The overall accuracy in function of the number of scales S in our multi-scale spectral feature extraction method.

fluctuates with S values above 36, but these fluctuations are within 2%. We choose $S = 36$ which yields nearly optimal performance on all the data sets. Larger values bring no significant benefit while they increase memory requirements.

We further analyze the effect of the proposed multi-scale spectral feature extraction (labelled by MSSFE) with different reduced versions: without using group convolution, without using shuffling operator, and without using squeeze and excitation (labelled by SE). The results in Fig. 11 show that the proposed MSSFE performs the best by combining these strategies, especially using the group convolution.

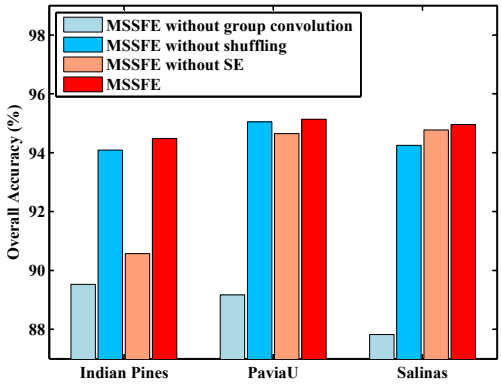

Fig. 11. The overall accuracy of the proposed multi-scale spectral feature extraction method (MSSFE) with different strategies for the three data sets.

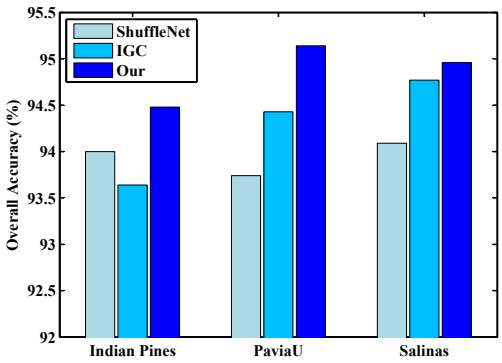

Fig. 12. The influence of different spectral-spatial extraction methods on the overall accuracy in three data sets. ShuffleNet [35] and IGC[36] are related group convolution-based methods.

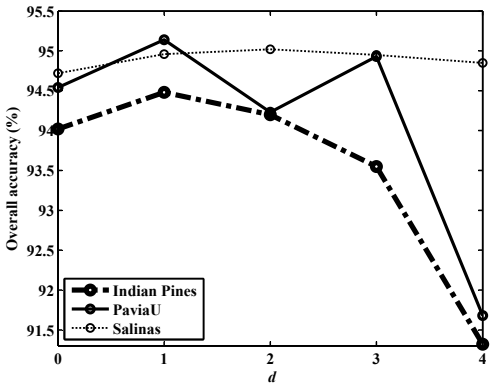

Fig. 13. The overall accuracy in function of the number of layers d in our spectral-spatial extraction stage on overall accuracy for the three data sets.

5) Analysis of the spectral-spatial feature extraction stage: We first compare the proposed discriminative spectral-spatial feature extraction approach with two state-of-the-art group convolution-based models: ShuffleNet [35], and IGC [36]. The results in Fig. 12 show clear advantage of our approach with the novel group competition structure compared to ShuffleNet and IGC for all the three data sets.

The results in Fig. 13 show the influence of the number of layers d on the performance of our discriminative feature extraction approach. $d = 1$ yields the best accuracy for all the three data sets. Observe that the case $d = 0$ corresponds to excluding the middle discriminative feature extraction stage. The results show clearly the benefit of this middle stage and indicate that its single-layer implementation is not only the simplest computationally but also optimal in terms of the overall network performance.

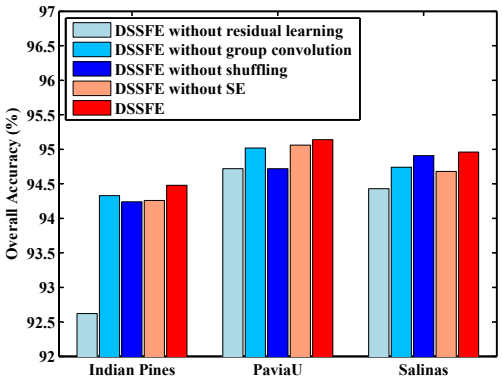

Fig. 14. The overall accuracy of the proposed discriminative spectral-spatial feature extraction method (DSSFE) with different strategies in three data sets.

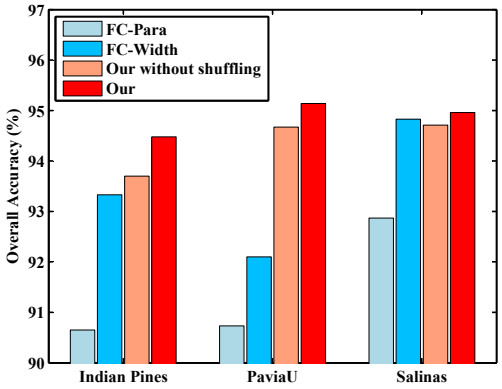

Fig. 15. The effect of different fusion methods on overall classification accuracy in three data sets. Our approach is compared to its version without shuffling and to fusion with fully connected layers with the same number of parameters as in our method (FC-Para) and with the same network width as in our method (FC-Width).

We further compare the proposed discriminative spectral-spatial feature extraction method (labelled by DSSFE) with its reduced versions: without using residual learning, without using group convolution, without using shuffling operator, and without using SE under the same setting as in Table II. The results in Fig. 14 show that employing these strategies indeed improves the overall accuracy for all the three data sets, especially employing the residual learning strategy.

6) Analysis of the feature fusion stage: We compare the performance of the proposed feature fusion method against fully connected layers with the same network width and with the same number of parameters and also to group fusion without using shuffling operation. The results in Fig. 15 show that the proposed lightweight fusion method consistently yields better accuracy than the versions with fully connected layers. It is also evident that the shuffling operation indeed improves the classification accuracy, especially in Indian Pines and PaviaU images. Furthermore, our analysis shows that two

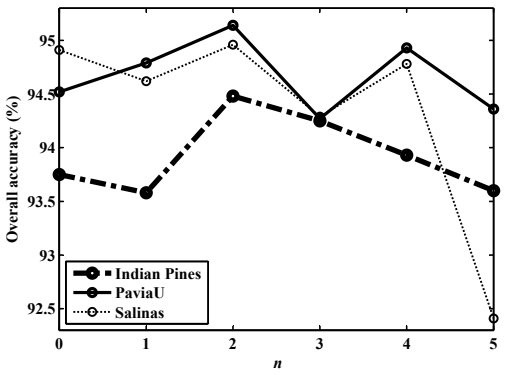

Fig. 16. The effect of the numbers of fusion layers n on overall classification accuracy in three data sets.

TABLE VII
COMPARISON OF THE COMPUTATIONAL COMPLEXITY ON DIFFERENT MODELS FOR THE PAVIAU IMAGE.

Method	Training (s)	Testing (s)	# Params ($\times$ M)	FLOPs ($\times 10^7$)
DRCNN	243.8	45.8	2.922	3.436
SSRN	205.5	17.1	0.198	2.858
HybridSN	10.4	6.3	0.741	2.983
ADGAN	87.9	3.9	7.382	71.805
MCNNCP	9.6	2.6	1.377	2.917
GCNN	89.2	6.9	27.548	5.549
FGCNN	171.4	7.5	3.077	0.655

shuffling layers are sufficient. The results in Fig. 16 show that two shuffling layers provide the best overall accuracy for the three data sets.

D. Computational Complexity

The results in Table VII provide comparative analysis of the computational complexity of the proposed FGCNN, our earlier conference version (GCNN) [43], and five representative reference methods: DRCNN [49], SSRN [44], HybridSN [54], ADGAN [17], and MCNNCP [42]. The following attributes are reported: the training and testing time, the number of parameters and the number of floating point operations (FLOPs). The reported values correspond to one of the data sets (PaviaU) and are similar for other two test data sets. All experiments are conducted on a computer equipped with an Intel Core i7-7820X CUP with 3.6 GHz and an Nvidia TITAN Xp GPU.

In the training and testing processes, our FGCNN is moderately fast compared to the reference methods. SSRN and HybridSN involve much less parameters than our method, which is in this respect comparable to DRCNN and MCNNCP, and much better than ADGAN and GCNN. Our FGCNN requires much less FLOPs compared to all the reference methods. Given that the accuracy of the proposed method is much better compared to the faster methods, it can be concluded that the proposed FGCNN with the same hyperparameter settings is very competitive and robust in terms of the classification accuracy compared to the current state-of-the-art.

V. CONCLUSION

In this paper, we proposed a fully group CNN architecture for robust spectral-spatial classification of hyperspectral images. It offers a unified theoretical framework where all the stages of the learning architecture are formulated as cascades of shuffled group convolutions. One of the key contributions is an original multi-scale spectral feature extraction approach. Its core component is a multi-kernel depthwise convolution that extends the regular depthwise convolution in order to weight the information from multiple scales. Within the same unified framework, we designed a discriminative spectral-spatial feature extraction approach with a novel group competition block to extract more discriminative spectral-spatial features with fewer learning parameters. Finally, our lightweight feature fusion stage sharply reduces the fusion parameters while enhancing the feature fusion capability. Experimental results on real data demonstrated robust performance compared to the current state-of-the-art.

REFERENCES

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, 2013.

[2] P. Ghamisi, E. Maggiori, S. Li, R. Souza, Y. Tarablaka, G. Moser, A. De Giorgi, L. Fang, Y. Chen, M. Chi *et al.*, "New frontiers in spectral-spatial hyperspectral image classification: the latest advances based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 3, pp. 10–43, 2018.

[3] G. Shi, H. Huang, and L. Wang, "Unsupervised dimensionality reduction for hyperspectral imagery via local geometric structure feature learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1425–1429, 2020.

[4] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017.

[5] 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, 2019.

[6] H. Huang, C. Pu, Y. Li, and Y. Duan, "Adaptive residual convolutional neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2520–2531, 2020.

[7] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, 2019.

[8] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, 2017.

[9] 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, 2019.

[10] 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, 2016.

[11] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classi-

fication," *IEEE Trans. Geosci. Remote Sens.*, early access, Aug. 18, 2020, doi: 10.1109/TGRS.2020.3015157.

[12] 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, 2020.

[13] L. Mou, X. Lu, X. Li, and X. X. Zhu, "Nonlocal graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8246–8257, 2020.

[14] X. Li, M. Ding, and A. Pižurica, "Deep feature fusion via two-stream convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2615–2629, 2020.

[15] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, 2019.

[16] X. He, Y. Chen, and P. Ghamisi, "Dual graph convolutional network for hyperspectral image classification with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, early access, Mar. 08, 2021, doi: 10.1109/TGRS.2021.3061088.

[17] J. Wang, F. Gao, J. Dong, and Q. Du, "Adaptive dropblock-enhanced generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Aug. 21, 2020, doi: 10.1109/TGRS.2020.3015843.

[18] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, 2018.

[19] 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, 2021.

[20] M. Denil, B. Shakibi, L. Dinh, N. De Freitas *et al.*, "Predicting parameters in deep learning," in *Adv. Neural. Inf. Process. Syst.*, 2013, pp. 2148–2156.

[21] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *Proc. ICML*. PMLR, 2020, pp. 8093–8104.

[22] 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, 2017.

[23] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, 2016.

[24] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "DLA-MatchNet for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Nov. 06, 2020, doi: 10.1109/TGRS.2020.3033336.

[25] Z. Li, M. Liu, Y. Chen, Y. Xu, W. Li, and Q. Du, "Deep cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Feb. 17, 2021, doi: 10.1109/TGRS.2021.3057066.

[26] J. Feng, H. Yu, L. Wang, X. Cao, X. Zhang, and L. Jiao, "Classification of hyperspectral images based on multiclass spatial–spectral generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5329–5343, 2019.

[27] S. Pande, B. Banerjee, and A. Pižurica, "Class reconstruction driven adversarial domain adaptation for hyperspectral image classification," in *Proc. IbPRIA 2019*. Springer, 2019, pp. 472–484.

[28] Y. Chen, K. Zhu, L. Zhu, X. He, P. Ghamisi, and J. A. Benediktsson, "Automatic design of convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7048–7066, 2019.

[29] J. Wang, R. Huang, S. Guo, L. Li, M. Zhu, S. Yang, and L. Jiao, "NAS-guided lightweight multiscale attention fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Jan. 21, 2021, doi: 10.1109/TGRS.2021.3049377.

[30] Y. Chen, L. Zhu, P. Ghamisi, X. Jia, G. Li, and L. Tang, "Hyperspectral images classification with Gabor filtering and convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2355–2359, 2017.

[31] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, 2018.

[32] F. I. Alam, J. Zhou, A. W.-C. Liew, X. Jia, J. Chanussot, and Y. Gao, "Conditional random field and deep feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1612–1628, 2019.

[33] X. Zhao, R. Tao, W. Li, H. C. Li, Q. Du, W. Liao, and W. Philips, "Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7355–7370, 2020.

[34] A. Santara, K. Mani, P. Hatwar, A. Singh, A. Garg, K. Padia, and P. Mitra, "BASS Net: Band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5293–5301, 2017.

[35] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2018, pp. 6848–6856.

[36] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2017, pp. 4373–4382.

[37] G. Xie, J. Wang, T. Zhang, J. Lai, R. Hong, and G.-J. Qi, "Interleaved structured sparse convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2018, pp. 8847–8856.

[38] K. Sun, M. Li, D. Liu, and J. Wang, "IGCV3: Interleaved low-rank group convolutions for efficient deep neural networks," *CoRR*, vol. abs/1806.00178, 2018. [Online]. Available: http://arxiv.org/abs/1806.00178

[39] X. Wang, M. Kan, S. Shan, and X. Chen, "Fully learnable group convolution for acceleration of deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2019, pp. 9041–9050.

[40] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, no. 99, pp. 1–17, 2018.

[41] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active learning with convolutional neural networks for hyperspectral image classification using a new Bayesian approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6440–6461, 2018.

[42] J. Zheng, Y. Feng, C. Bai, and J. Zhang, "Hyperspectral image classification using mixed convolutions and covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 522–534, 2021.

[43] X. Li, M. Ding, and A. Pižurica, "Group convolutional neural networks for hyperspectral image classification," in *Proc. 2019 IEEE Int Conf on Image Process (ICIP)*, 2019, pp. 639–643.

[44] 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, 2018.

[45] 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, 2017.

[46] Z. Wang, C. Zou, and W. Cai, "Small sample classification of hyperspectral remote sensing images based on sequential joint deeping learning model," *IEEE Access*, vol. 8, pp. 71 353–71 363, 2020.

[47] X. Zhou, S. Li, F. Tang, K. Qin, S. Hu, and S. Liu, "Deep learning with grouped features for spatial spectral classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 97–101, 2017.
[48] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2017, pp. 1251–1258.
[49] 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, 2018.
[50] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2018, pp. 4510–4520.
[51] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.
[52] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger, "CondenseNet: An efficient densenet using learned group convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2018, pp. 2752–2761.
[53] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.
[54] 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, 2020.
[55] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, and R. Wang, "Deep few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2290–2304, 2019.

Xian Li (S'19) received the M.S. degree from Harbin Institute of Technology, Harbin, China, in 2016, where he is currently pursuing the Ph.D. degree in instrument science and technology with the School of Instrumentation Science and Engineering. From 2018 to 2020, he was a doctoral researcher with the Department of Telecommunications and Information Processing, UGent-GAIM, Ghent University, Belgium, supported by the China Scholarship Council. His research interests include deep learning, hyperspectral remote sensing image analysis.

Mingli Ding received the B.S. and the Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 2000 and 2005, respectively. From 2009 to 2010, he was a Visiting Scholar with the French National Center for Scientific Research, Toulouse, France. He is currently a Full Professor with the School of Instrumentation Science and Engineering, Harbin Institute of Technology, China.

His research interests include deep learning, image classification, object detection, automation test technology, and information processing.

Aleksandra Pižurica (SM'15) received the Diploma in electrical engineering from the University of Novi Sad, Serbia, in 1994, the Master of Science degree in telecommunications from the University of Belgrade, Serbia, in 1997, and the Ph.D. degree in engineering from Ghent University, Belgium, in 2002.

She is a Professor in statistical image modeling with Ghent University. Her research interests include the area of signal and image processing and machine learning, including multiresolution statistical image models, Markov Random Field models, sparse coding, representation learning, and image and video reconstruction, restoration, and analysis.

Prof. Pižurica served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (2012 – 2016), Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (2016 – 2019) and currently an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. She was also the Lead Guest Editor for the EURASIP Journal on Advances in Signal Processing for the Special Issue "Advanced Statistical Tools for Enhanced Quality Digital Imaging with Realistic Capture Models" (2013). The work of her team has been awarded twice the Best Paper Award of the IEEE Geoscience and Remote Sensing Society Data Fusion contest, in 2013 and 2014. She received the scientific prize "de Boelpaepe" for 2013–2014, awarded by the Royal Academy of Science, Letters and Fine Arts of Belgium for her contributions to statistical image modeling and applications to digital painting analysis.