GROUP CONVOLUTIONAL NEURAL NETWORKS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Xian Li[1,2], *Mingli Ding*[1] *and Aleksandra Pižurica*[2]

[1]School of Instrumentation Science and Engineering, Harbin Institute of Technology, China
[2]Department of Telecommunications and Information Processing,
Ghent University-imec, Belgium

ABSTRACT

Convolutional Neural Network (CNN) has been widely applied in hyperspectral image (HSI) classification exhibiting excellent performance. The CNN model overfitting is a common issue in this domain due to limited amount of labelled training samples. In addition, making the full use of spectral information is still considered an open problem. In this paper, we propose a novel group 2D-CNN model for spectral-spatial classification. Specifically, we propose an original multi-scale spectral feature extraction approach based on a novel concept of multi-kernel depthwise convolution. Furthermore, we exploit for the first time shuffle operation on the group convolutions in HSI spectral-spatial feature extraction to effectively limit the amount of learning parameters. As a result, we design a small and efficient network for HSI classification. Experimental results on real data demonstrate favourable performance compared to the current state-of-the-art.

***Index Terms*—** Group convolutional neural networks, multi-scale spectral feature extraction, hyperspectral image.

1. INTRODUCTION

Compared with panchromatic and multi-spectral remote sensing images, hyperspectral image (HSI) contains much richer spectral information, which enables more accurate discrimination between different materials or objects in the target scene [1]. However, this large dimensionality in the spectral domain poses huge challenges for processing due to scarcity of labelled training samples [2]. Feature extraction is typically employed to alleviate this problem [3].

Recent studies demonstrate huge potential of deep learning, and in particular, convolutional neural network (CNN) architectures for feature extraction in HSI [4, 5]. CNNs with one dimensional neurons (1D-CNN) are used to extract spectral features and reportedly achieve better performance than traditional machine learning methods [6].

In general, making use of the spatial context together with spectral information yields better classification performance than spectral classification alone. Spectral-spatial feature extraction methods based on 2D-CNN and 3D-CNN were reported, e.g., in [4, 7]. 2D-CNN based models [4, 8] often exploit the first few principal components of the HSI bands to extract the spatial features. This way, fewer learning parameters are needed, but the spectral information is less well exploited [9]. An alternative approach based on dual-channel CNN solves this problem by extracting the spectral and spatial features separately [10, 11]. Yue *et al.* [12] presented a deep learning framework where the spectral features and spatial features were extracted via stacked auto-encoders (SAE) and CNN. Hao *et al.* [13] proposed an improved and more robust version by using stacked denosing autoencoders (SdAE), which are more robust to noise than SAE and achieve better performance. Recently, a unified framework based on long short-term memory (LSTM) model and CNN was reported to extract spectral and spatial features in [14].

An alternative approach based on 3D-CNN models extracts spectral-spatial features simultaneously without dimensionality reduction. E.g., the 3D-CNN model in [4, 9] extracts the integrated spectral-spatial features as the structural characteristics of 3D HSI. Lately, a residual learning version was presented in [15]. However, 3D-CNN models often employ smaller networks than 2D-CNN to avoid overfitting, so it is hard to extract deep features and fully use the global spectral information. Consequently, the classification map tends to be oversmoothed [14].

A major challenge faced by all the methods mentioned above is how to avoid overfitting under a limited amount of labelled training samples. To mitigate this problem, group features extraction methods were recently proposed in [16, 17]. A downside is a weaker representation due to ignoring the correlation among the different groups. A clever idea of channel shuffling, dubbed ShuffleNet, was recently introduced in computer vision [18, 19], specifically for mobile devices, to overcome the limitations of group convolutions.

In this paper, we investigate a different perspective of optimizing the network structure while enhancing the efficiency of feature extraction. We propose a novel group 2D-CNN model for HSI spectral and spatial classification. While most

This work was supported by the China Scholarship Council and the Research Foundation Flanders (FWO) under the grant G.OA26.17N.

of 2D-CNN based methods like [4, 8, 11–14] include some kind of dimensionality reduction (such as PCA), we introduce instead a multi-scale spectral feature extraction (MSSFE) module in the first convolution layer, without any dimensionality reduction. The core of our proposed MSSFE module is a novel concept of multi-kernel depthwise convolution that we formulate in order to weight the spatial information of each band from different scales. We combine this multi-scale representation with group convolution to fully extract global spectral features. Then we employ two group convolution layers with shuffle operation to efficiently extract group spectral and spatial features under less learning parameters. As a result, we design a small and efficient network to extract spectral and spatial features of HSI.

The main contributions of this paper are:

1) We introduce multi-kernel depthwise convolution to weight the spatial information at different scales in each band.

2) Based on this concept, we develop a novel MSSFE module, which effectively extracts multi-scale global spectral information making use of *all* spectral bands.

3) We design an architecture that exploits group convolutions with shuffle operation to extract spectral-spatial features effectively. To our knowledge, the use of shuffle operation for HSI group feature extraction has not been reported before.

The rest of this paper is organized as follows. Section 2 introduces the proposed method. The experiments are reported in Section 3 and Section 4 concludes the paper.

2. PROPOSED METHOD

2.1. Overall architecture

A major challenge faced by CNN-based models for HSI classification is overtraining because of insufficient amount of labelled training samples to justify the model parameters. How to fully extract spectral features and effectively fuse them with the spatial features is another core problem that still requires further research. In order to address these challenges, we propose a novel group 2D-CNN architecture to extract and fuse group spectral and spatial features effectively. Fig. 1 shows the overall architecture of the proposed method, which consists of three parts: 1) multi-scale spectral feature extraction, 2) group spectral and spatial feature extraction, and 3) grouped feature fusion and classification. Specifically, we propose a novel MSSFE module to extract multi-scale global spectral information in the beginning stage. Then, we exploit the group convolutions with shuffle operation to extract group spectral and spatial features. Finally, two fully connected layers fuse the extracted features, and the label of each pixel is predicted by a softmax layer.

2.2. Multi-scale spectral feature extraction module

Since HSI contain hundreds of bands, current 2D-CNN based models often use only few principle components as inputs to reduce the amount of the parameters [8–14], sacrificing some useful information for the classification [17]. On the other hand, the data-driven feature learning methods that better preserve the useful information and fully extract spectral features, suffer from the phenomena known as curse of dimensionality [2]. To mitigate this problem, we propose a novel multi-scale spectral feature extraction approach, termed MSSFE, which boosts the spectral feature extraction power under an acceptable number of learning parameters.

The core component of the proposed MSSFE is a novel multi-kernel depthwise convolution operation. Different from the common depthwise convolution [20], it has multiple output channels for each input channel, so we refer to it as multi-kernel depthwise convolution. Mathematically, with a depthwise convolution, an input neuron at location (i, j) of the k-th feature map in the $(l-1)$-th layer leads to a single output:

$$x_{i,j}^{l,k} = \sigma\Big(\sum_{p=0}^{H_l-1}\sum_{q=0}^{W_l-1} w_{pq}^{l,k} \cdot x_{(i+p)(j+q)}^{(l-1)k} + b^{l,k}\Big). \qquad (1)$$

where $\cdot$ denotes the element-wise product, and H_l and W_l are the height and the width of the kernels. $w_{pq}^{l,k}$ is the weight and $b^{l,k}$ is the bias term. σ refers to the activation function. We now define a multi-kernel depthwise convolution with s scales, where the corresponding input neuron yields s outputs: $(x_{i,j}^{l,k,1}, ..., x_{i,j}^{l,k,s})$, and the n-th output $x_{i,j}^{l,k,n}$ is

$$x_{i,j}^{l,k,n} = \sigma\Big(\sum_{p=0}^{H_l-1}\sum_{q=0}^{W_l-1} w_{npq}^{l,k} \cdot x_{(i+p)(j+q)}^{(l-1)} + b_n^{l,k}\Big). \qquad (2)$$

Now, the weights $w_{npq}^{l,k}, 1 \leq n \leq s$, are learned for each scale. We employ 1×1 multi-kernel depthwise convolution to weight the spatial information of each band from s scales as shown in Fig. 1. In Fig. 2, we show a more detailed structure of MSSFE. With the shuffle operation, we transform the weighted HSI into s groups, where each group contains all the bands of the HSI at a particular scale. Each shuffled group is then fed to the 1×1 convolution separately. The squeeze and excitation (SE) [21] module emphasizes informative bands (e.g., features) and suppresses useless ones (e.g., noisy).

2.3. Group spectral and spatial feature extraction module

The powerful feature extraction ability of CNN results from a large number of filters. This translates to a large number of learning parameters. However, the amounts of labelled training samples for HSI are rather limited, and thus current CNN-based models tend to be overfitting. Group features extraction methods mitigate this problem [16, 17], but they may suffer from some performance loss because of ignoring the correlation among the different groups. Inspired by ShuffleNet [18, 19], we propose two group convolution layers with channel shuffle operation to replace two regular convolution layers. Mathematically, the value of a neuron at location (i, j) of

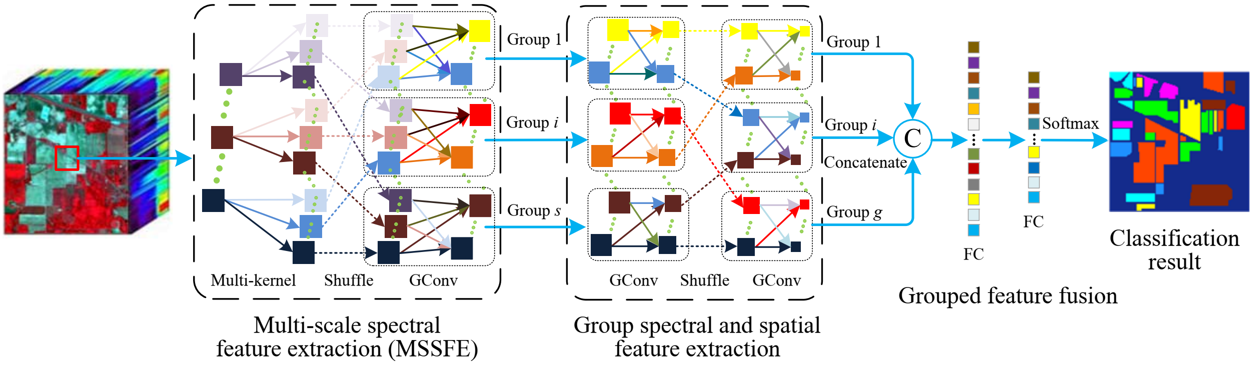


Fig. 1. The architecture of the proposed method. Different colors of the arrows illustrate here different weights.

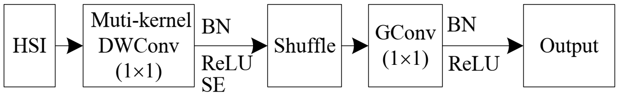


Fig. 2. A detailed structure of MSSFE.

the k-th feature map in the l-th layer of a group convolution with g groups is

$$x_{i,j}^{l,k} = \sigma\Big(\sum_{m=1}^{M/g}\sum_{p=0}^{H_l-1}\sum_{q=0}^{W_l-1} w_{pq}^{l,k,m}\cdot x_{(i+p)(j+q)}^{(l-1)m} + b^{l,k}\Big). \qquad (3)$$

where m indexes the input feature map. Compared with a regular convolution layer, a group layer reduces the number of parameters g times and mitigates this way the overfitting. Let $\mathbf{F}_g = [(f_{1,1}, ..., f_{1,M/g}), ..., (f_{i,1}, ..., f_{i,M/g}), ..., (f_{g,1}, ..., f_{g,M/g})]$ denote the outputs of the first group layer, and $(f_{i,1}, ..., f_{i,M/g})$ denote all the M/g elements in the i-th group. Without shuffle operation, the output of each group in the second group layer is only derived from the corresponding local input (i.e., the i-th output is derived from $(f_{i,1}, ..., f_{i,M/g})$) resulting in a weak representation. Shuffle operation transforms $\mathbf{F}_g$ into $\mathbf{F}'_g = [(f_{1,1}, ..., f_{i,1}, ..., f_{g,1}), ..., (f_{1,M/g}, ..., f_{i,M/g}, ..., f_{g,M/g})]$. Now, the output of each group in the second convolution layer comes from different input groups. The shuffle operation establishes hereby the correlation among the groups, and thus we can extract group spectral and spatial features more effectively, especially when the labelled data are very limited.

2.4. Grouped feature fusion and classification

We exploit two fully connected layers to concatenate and fuse all the extracted features. To avoid overfitting, we use a dropout with 0.2 threshold before the grouped feature fusion. Finally, we employ a softmax layer with L2 regularization, and express the cost as

$$c = \mathcal{L}(\mathbf{P}, \mathbf{Y}, \mathbf{W}) + \lambda\|\mathbf{W}\|_{\mathrm{F}}^2. \qquad (4)$$

where $\mathcal{L}$ denotes the cross entropy loss function, $\mathbf{P}$, $\mathbf{Y}$, $\mathbf{W}$ are the predicted labels, the ground truth, and the fusion weights, respectively. $\| \quad \|_F^2$ is the Frobenius norm and λ is a regularization parameter. We optimize (4) by using the mini-batch adadelta [22].

3. EXPERIMENTAL RESULTS AND ANALYSIS

Experiments are conducted on two real HSIs as detailed later. We compare our proposed method with the following state-of-the-art CNN-based methods: CNN [6], PPFCNN [23], CDCNN [24], and SSCNN [8]. The parameters of these methods are set to the default values indicated in their original works. The overall accuracy (OA) and average accuracy (AA) are used for quantitative evaluation. The experiments are repeated ten times by randomly selecting 200 training samples per class and 10% of them as validation samples, and the average over the ten runs is reported. The patch size in our method is set to 5×5. To accelerate the training process, the ReduceLROnPlateau and EarlyStoping functions are adopted. The base learning rate is set to 1 and 3 for Indian Pines and PaviaU images, respectively. We set the epochs and batch size as 500 and 64, respectively.

3.1. Experiments on real data

Experiment 1 was conducted on the *Indian Pines*, which was captured by the Airborne/Visible Infrared Imaging Spectrometer sensors from the North-western Indiana in June 1992. It contains 16 classes, out of which we select 8 large classes. 4 water absorption bands were removed. The results in Table 1 show a significant improvement of the proposed method over the reference methods. In comparison with CNN, PPFCNN, CDCNN, and SSCNN, the increase in OA is 11.74%, 4.85%, 4.51% and 2.12%, respectively.

Experiment 2 was conducted on an urban HSI: *University of Pavia* (denoted as PaviaU), which was acquired by the Reflective Optics System Imaging Spectrometer sensor during a flight campaign over Pavia, Northern Italy. It consists of

Table 1. Classification accuracy for *Indian Pines*.

Classes	CNN	PPFCNN	CDCNN	SSCNN	Proposed
1	78.58	92.99	90.1	96.28	97.70±0.75
2	85.24	96.66	97.1	92.26	99.43±0.50
3	96.10	98.58	100	99.30	99.86±0.24
4	99.64	100	100	100	100±0
5	89.64	96.24	95.9	92.84	98.91±0.74
6	81.55	87.80	87.1	98.21	98.02±0.61
7	95.42	98.98	96.4	92.45	99.90±0.24
8	98.59	99.81	99.4	98.98	99.94±0.05
OA(%)	87.01	93.90	94.24	96.63	98.75±0.24
AA(%)	90.60	96.38	95.75	96.29	99.22±0.14

Table 2. Classification accuracy for *PaviaU*.

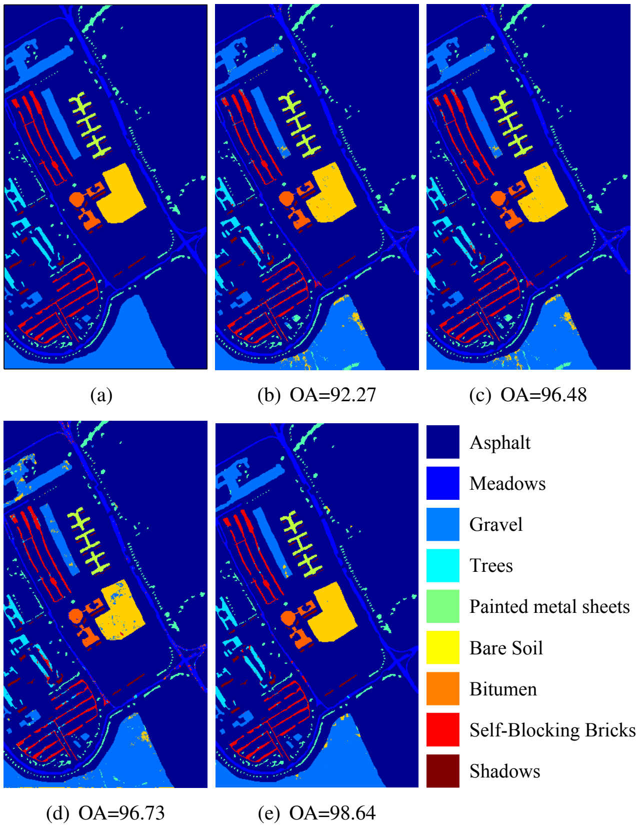
Classes	CNN	PPFCNN	CDCNN	SSCNN	Proposed
1	88.38	97.42	94.6	97.40	98.26±0.44
2	91.27	95.76	96	99.40	98.97±0.26
3	85.88	94.05	95.5	94.84	95.58±1.19
4	97.24	97.52	95.9	99.16	99.10±0.41
5	99.91	100	100	100	100±0
6	96.41	99.13	94.1	98.70	99.66±0.27
7	93.62	96.19	97.5	100	99.53±0.38
8	87.45	93.62	88.8	94.57	96.29±1.43
9	99.57	99.60	99.5	99.87	99.97±0.06
OA(%)	92.27	96.48	96.73	98.41	98.63±0.19
AA(%)	93.36	97.03	95.77	98.22	98.60±0.24

Table 3. The effect of the number of scales on OA.

Datasets	RConv1	s=10	s=20	s=30	s=40
Indian Pines	95.84	98.28	98.50	98.75	98.51
PaviaU	97.98	98.50	98.56	98.63	98.51

Table 4. The effect of the group convolutions and shuffle operation on OA.

Datasets	AllRConv	RConv2	NoShuffle	Shuffle(g=10)
Indian Pines	94.93	98.41	98.63	98.75
PaviaU	97.82	98.48	98.57	98.63


(a) (b) OA=92.27 (c) OA=96.48

(d) OA=96.73 (e) OA=98.64

Fig. 3. Classification results on *PaviaU*. (a) Ground truth, and classification maps of (b) CNN, (c) PPFCNN, (d) CDCNN, and (e) the proposed method.

610×340 pixels with 103 spectral bands and 9 classes. The ground truth is shown in Fig. 3(a). The results on PaviaU are reported in Table 2 and Fig. 3. Our proposed method consistently yields better accuracy than the other four methods. Visually, our method presents more accurate and more similar results to the reference map. This can be observed clearly e.g., in the regions of Meadows and Bare Soil.

In addition, we verify the effectiveness of the MSSFE module. Table 3 lists the results with a regular convolution layer (denoted as RConv1) versus the MSSFE module with different number of scales s on the two real HSIs. Clearly, the OA value with MSSFE is better than with regular convolution layer, which demonstrates the improved spectral feature extraction ability. Table 4 reports comparative results for the regular convolution model (denoted as AllRConv), two regular convolution layers (denoted as RConv2), two group convolutional layers without shuffle operation (denoted as NoShuffle) and with shuffle operation with $g = 10$ groups. The results verify that the group convolutions with shuffle operation indeed improve the accuracy due to avoiding the overfitting and considering the correlation among the different groups.

4. CONCLUSION

In this paper, we propose a novel group 2D-CNN architecture for HSI spectral and spatial classification. In the first convolution layer, we propose a novel MSSFE module to efficiently extract global spectral features. The core of this module is a multi-kernel depthwise convolution that we defined by extending the standard depthwise convolution, in order to weight the spatial information from multiple scales. We combine shuffle operation with group convolution to extract HSI group spectral and spatial features with less learning parameters. Experimental results on real data demonstrate favourable performance compared to the current state-of-the-art.

5. REFERENCES

[1] David Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 17–28, 2002.

[2] Gordon Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE transactions on information theory*, vol. 14, no. 1, pp. 55–63, 1968.

[3] José M Bioucas-Dias, Antonio Plaza, Gustavo Camps-Valls, Paul Scheunders, Nasser Nasrabadi, and Jocelyn Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[4] Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[5] Pedram Ghamisi, Emmanuel Maggiori, Shutao Li, Roberto Souza, Yuliya Tarablaka, Gabriele Moser, Andrea De Giorgi, Leyuan Fang, Yushi Chen, Mingmin 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 Geoscience and Remote Sensing Magazine*, vol. 6, no. 3, pp. 10–43, 2018.

[6] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, 2015.

[7] Mathieu Fauvel, Yuliya Tarabalka, Jon Atli Benediktsson, Jocelyn Chanussot, and James C Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, 2013.

[8] Shaohui Mei, Jingyu Ji, Junhui Hou, Xu Li, and Qian Du, "Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4520–4533, 2017.

[9] Ying Li, Haokui Zhang, and Qiang Shen, "Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network," *Remote Sensing*, vol. 9, no. 1, pp. 67, 2017.

[10] Haokui Zhang, Ying Li, Yuzhu Zhang, and Qiang Shen, "Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network," *Remote Sensing Letters*, vol. 8, no. 5, pp. 438–447, 2017.

[11] Jingxiang Yang, Yong-Qiang Zhao, and Jonathan Cheung-Wai Chan, "Learning and transferring deep joint spectral–spatial features for hyperspectral classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4729–4742, 2017.

[12] Jun Yue, Shanjun Mao, and Mei Li, "A deep learning framework for hyperspectral image classification using spatial pyramid pooling," *Remote Sensing Letters*, vol. 7, no. 9, pp. 875–884, 2016.

[13] Siyuan Hao, Wei Wang, Yuanxin Ye, Tingyuan Nie, and Lorenzo Bruzzone, "Two-stream deep architecture for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2349–2361, 2018.

[14] Yonghao Xu, Liangpei Zhang, Bo Du, and Fan Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, , no. 99, pp. 1–17, 2018.

[15] Zilong Zhong, Jonathan Li, Zhiming Luo, and Michael Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 847–858, 2018.

[16] Xichuan Zhou, Shengli Li, Fang Tang, Kai Qin, Shengdong Hu, and Shujun Liu, "Deep learning with grouped features for spatial spectral classification of hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 1, pp. 97–101, 2017.

[17] Anirban Santara, Kaustubh Mani, Pranoot Hatwar, Ankit Singh, Ankur Garg, Kirti Padia, and Pabitra Mitra, "Bass net: band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5293–5301, 2017.

[18] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 6848–6856.

[19] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 116–131.

[20] Laurent Sifre and Stéphane Mallat, *Rigid-motion scattering for image classification*, Ph.D. thesis, Citeseer, 2014.

[21] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[22] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[23] Wei Li, Guodong Wu, Fan Zhang, and Qian Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 844–853, 2017.

[24] Hyungtae Lee and Heesung Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4843–4855, 2017.