Multi-branch Selective Kernel Networks for Hyperspectral Image Classification

T. Alipour-Fard, M. E. Paoletti, *Student Member, IEEE,* J. M. Haut, *Senior Member, IEEE,* H. Arefi, J. Plaza, *Senior Member, IEEE,* and A. Plaza, *Fellow, IEEE*

***Abstract*—Convolutional neural networks (CNNs) have demonstrated excellent performance in hyperspectral image (HSI) classification. However, tuning some critical hyper-parameters of a CNN –such as the receptive field (RF) size– presents a major challenge due to the presence of features with different scales in HSIs. Contrary to the conventional design of CNNs, which fizes the RF size, it has been proven that the RF size is modulated by the stimulus and hence depends on the scene being considered. Such a dilemma has been rarely considered in CNN design. In this letter, a new Multi-branch Selective Kernel Network (MSKNet) is introduced, in which the input image is convolved using different RF sizes to create multiple branches, so that the effect of each branch is adjusted by an attention mechanism according to the input contrast. As a result, our newly developed MSKNet is capable of modeling different scales. Our experimental results, conducted on three widely-used HSIs, reveal that the MSKNet can outperform state-of-the-art CNNs in the context of HSI classification problems. The source code of our newly developed MSKNet is available from: https://github.com/mhaut/MSKNet-HSI**

***Index Terms*—deep learning, hyperspectral images (HSIs), convolutional neural networks (CNNs), receptive field (RF), selective kernel networks (SKNets).**

I. Introduction

HYPERSPECTRAL images (HSIs) comprise hundreds of images (at different wavelength channels) for the same area on the surface of the Earth. The resulting datacubes provide an opportunity to detect and recognize a wide variety of objects. Classification is one of the most important techniques to extract information from HSIs. However, there are some general challenges for the successful classification of HSIs, including the high dimensionality of the data, the limited availability of training samples (which hampers supervised classification techniques), or the correlation between spectral signatures belonging to different classes. To address these problems, spatial information has been used as a complement to spectral information in HSI classification [1]. In a recent study [2], by considering features in spatial and frequency domains, invariant attribute profiles were adopted to address the different semantic characteristics of the input patches with the same centered pixel label. Convolutional neural networks (CNNs) have become the state-of-the-art of supervised techniques, due to their ability to perform automatic feature generation and also to their generalization power [3]. Despite the advantages of CNNs, the lack of training data and the large number of hyper-parameters involved in the training of CNNs (with the subsequent overfitting problem) have become important obstacles to CNN-based HSI classification.

Among the CNNs hyper-parameters, the RF plays a very important role [4], [5]. The RF is the region of the input space that affects a particular unit of the CNN. Covering different RFs is important to recognize features with different scales and sizes at a specific layer of the CNN [4]. Fixing the RF size on CNNs is an inefficient assumption. This is because the visual cortex has the ability to collect information with different scales at the same processing level [6], [7]. If the RF size is selected to be too large, it can eliminate fine-grained structures. If the RF size is selected to be too small, it can remove coarse-grained structures. In both cases, the HSI classification accuracy can be significantly reduced.

To overcome the disadvantage of using single-branch CNNs, solutions for achieving an optimal architecture have been developed in the computer vision literature [8], [9], [10]. Specifically, multi-branch approaches were introduced to create branches with different RF sizes and combine them to obtain highly informative feature maps.The GoogleNet [8] incorporated an inception module, in which different branches were generated by different RF sizes to aggregate/concatenate information from different scales. The main weakness of GoogleNet lies in the fact that its linear aggregation approach may be insufficient to provide a powerful combination strategy [5]. In addition, various methods such as grouped/depth-wise/dilated convolutions have been introduced in order to reduce the number of hyper-parameters while incorporating parallel processing strategies [11]. Moreover, Hang et al. in [12] designed a two-layer cascade recurrent neural network (where the first layer removes redundant information and the second layer learns in complementary fashion). Afterward, the optimal weights for the fusion of the features from the two layers are calculated through a gated recurrent unit.

Another relevant development along the aforementioned lines is the highway network architecture presented in [13], which uses a gating mechanism to modulate the flow of information from different branches and create a deep network. The training of the highway network is difficult, mainly because of the gradient vanishing problem that resulted in

This work has been supported by the Spanish Ministry (FPU15/02090), Junta de Extremadura, Ref. GR18060 and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 734541 (EOXPOSURE). (Corresponding author: Juan M. Haut.)

T. Alipour-Fard and H. Arefi are with the School of Surveying and Geospatial Engineering, University of Tehran, Tehran, Iran. (e-mail: tayebalipour@gmail.com; hossein.arefi@ut.ac.ir).

M. E. Paoletti, J. M. Haut, J. Plaza and A. Plaza are with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain.(e-mail: mpaoletti@unex.es; juanmariohaut@unex.es; jplaza@unex.es; aplaza@unex.es).

the idea of ResNet (by utilizing skip connections) [9], [14]. FractalNet [15] and multi-level ResNet [16] methods were also designed to aggregate different branches recursively. Other than multi-branch methods, pyramidal structures (i.e., multi-scale systems) have also been considered. Specifically, Zhao et. al [17] proposed a multi-scale CNN (MCNN) to extract high-level spatial features for satellite image classification. In the MCNN, an image pyramid was constructed to capture spatial features across scales, and then high-level spatial features were combined with spectral features to train the CNN. Multi-scale covariance maps have also been proposed to extract hand-crafted features able to fully exploit the spectral-spatial information present on HSIs [18]. This was done by extracting patches with various sizes around the labeled pixels and then calculating the covariance matrix between the spectral bands. The authors in [19] presented a deformation-based convolution strategy for finding the optimal RF size for targets of interest in the image. A remaining challenge with multi-branch approaches is how to incorporate a mechanism to aggregate information from different branches in a non-linear manner. Recently, attention mechanisms have been developed to focus on key parts of the image, discarding irrelevant information [20] (see Fig. 1). Attention mechanisms can be used to recalibrate the feature response and to model adaptive, non-linear dependencies between feature maps with the gating mechanism.

In this letter, a new multi-branch selective kernel network (MSKNet) for HSI classication was developed. MSKNet incorporates an attention mechanism that conducts non-linear aggregation from different branches, addressing the inefficiency of the traditional (linear) aggregation approach by proposing an end-to-end framework that aggregates branch information by computing the contrast, and determining an effective weight by means of an attention mechanism. Our approach has been compared with a traditional CNN, using several HSI benchmark datasets. The experimental results demonstrate the superiority of the proposed approach, which outperforms CNNs with linear branch aggregation.

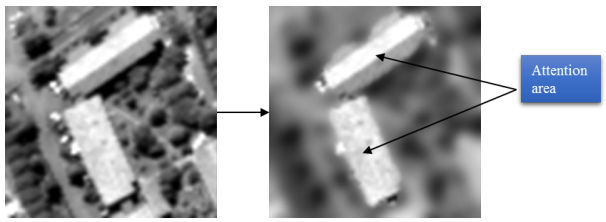

Fig. 1. An example illustrating the attention mechanism implemented by our MSKNet (the attention area is given by the two main buildings in the scene).

II. Methodology

Fig. 2 shows a general overview of the proposed MSKNet model for HSI classication, which is composed by three groups of selective kernel units (SKunits) followed by normalization-activation functions and a fully connected (FC) classifier at the end. The proposed workflow involves three main steps: (i) preprocessing of the HSI data cube, (ii) multi-kernel feature extraction, conducted by 2D convolutions, and (iii) selection of the most descriptive ones through an attention mechanism based on selective kernel layers.

A. Data Preprocessing

Let the HSI data cube be denoted by $\mathbf{X} \in \mathbb{R}^{H\times W\times B}$, where H is the height, W is the width and B is the number of spectral bands. $\mathbf{X}$ is split into training and testing sets, considering for each spectral pixel $\mathbf{x}_i$ a neighborhood window of size $S\times S\times B$, with the aim of providing spectral-spatial information to enhance the feature learning of the proposed model. These HSI patches are sent as input data to the proposed model, which is understood as a mapping function $\mathcal{M}_\theta : \mathbf{X} \rightarrow \mathbf{Y}$ that assigns for each $\mathbf{x}_i$ its corresponding land cover label y_i, obtaining the final classification map $\mathbf{Y} \in \mathbb{R}^{H\times W}$ by adjusting its learnable parameters θ.

B. Multi-Kernel Feature Extraction

The architectural body of the proposed model is composed by blocks of SKunits followed by normalization and non-linear activation functions. In this sense, SKunits implement a multi-branch architecture composed by depth-wise separable convolutional layers that exhibit different kernels size[1] $C\times k\times k\times C_{in}$ (depending on the branch in which they are located) that are intended to extract multi-kernel spectral-spatial features. Regarding to this, the convolutional-based transformations $\tilde{\mathcal{F}}^{(l)} : \mathbf{X}^{(l-1)} \rightarrow \tilde{\mathbf{U}}^{(l)}$ and $\hat{\mathcal{F}}^{(l)} : \mathbf{X}^{(l-1)} \rightarrow \hat{\mathbf{U}}^{(l)}$ (indicated as "split" step in Fig. 2) are applied to the original l-th SKunit's input (denoted as $\mathbf{X}^{(l-1)} \in \mathbb{R}^{S\times S\times C_{in}}$). These transformations apply 2D-grouped convolutions with kernels 3×3 and 5×5 respectively, adapting the zero-padding to maintain the spatial dimensions, and being followed by batch normalization (BN) and ReLU (Rectified Linear Unit) as non-linear activation function, resulting into the feature volumes $\tilde{\mathbf{U}}^{(l)} \in \mathbb{R}^{S\times S\times C}$ and $\hat{\mathbf{U}}^{(l)} \in \mathbb{R}^{S\times S\times C}$:

$$\begin{aligned}\tilde{\mathbf{U}}^{(l)} &= \tilde{\mathcal{F}}^{(l)}(\mathbf{X}^{(l-1)}) = \text{ReLU}\left(\beta\left(\mathbf{W}^{(l)} *_{C\times 3\times 3} \mathbf{X}^{(l-1)} + bias^{(l)}\right)\right)\\ \hat{\mathbf{U}}^{(l)} &= \hat{\mathcal{F}}^{(l)}(\mathbf{X}^{(l-1)}) = \text{ReLU}\left(\beta\left(\mathbf{W}^{(l)} *_{C\times 5\times 5} \mathbf{X}^{(l-1)} + bias^{(l)}\right)\right)\end{aligned} \quad (1)$$

where $*_{C\times k\times k}$ denotes the convolutional operation composed by C filters with receptive field $k\times k$, $\mathbf{W}^{(l)}$ and $bias^{(l)}$ are the weights and biases of the each convolutional layer that belongs to the l-SKunit, and β is the BN. Each element of $*_{C\times k\times k}$ operation is obtained as:

$$\dot{u}_{i,j}^{(l),c} = \sum_{\hat{c}=0}^{C_{in}-1}\sum_{\hat{i}=0}^{k-1}\sum_{\hat{j}=0}^{k-1}\left(w_{\hat{i},\hat{j},\hat{c}}^{(l)} \cdot x_{i+\hat{i},j+\hat{j},\hat{c}}^{(l-1)}\right) + bias^{(l)} \quad (2)$$

where $\dot{u}_{i,j}^{(l)}$ is the (i,j)-th element of the c-th feature map of volumes $\tilde{\mathbf{U}}^{(l)}$ or $\hat{\mathbf{U}}^{(l)}$, obtained at the l-th SKunit.

[1]Convolutional layers are defined by an n-dimensional kernel $C_{out}\times k\times k\times C_{in}$, where C_{in} is the number of feature maps from the input volume and C_{out} is the number of filters with receptive field $k\times k$.

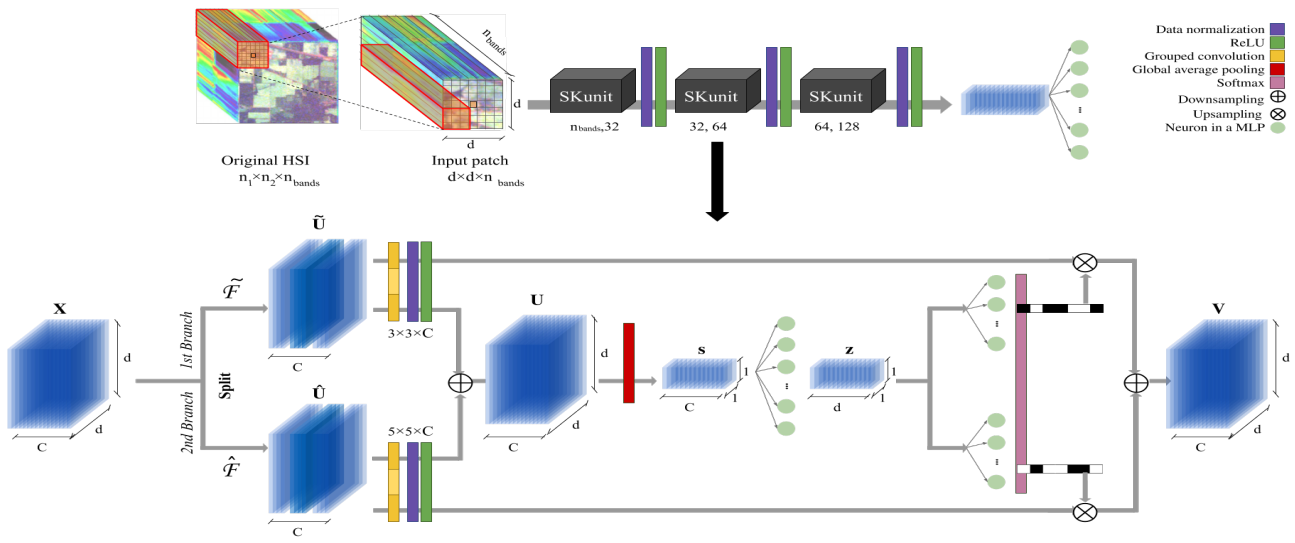

Fig. 2. General flowchart of the proposed multi-branch selective kernel network (MSKNet) and the selective kernel building block (SKunit) for hyperspectral image (HSI) classification. Usually, we set $C = C_{in}$ and $r = 2$.

C. Selective Kernel Attention Mechanism

Once the multi-kernel feature extraction is performed, control gates are required to regulate the information flow between the different branches, with the aim of combining and enhancing the most descriptive features in the subsequent layers. The final goal is to allow our model to adaptively adjust the receptive field $k \times k$ to handle different scales of information. The selective kernel strategy (based on attention mechanism) has been implemented in two main steps: multi-kernel data fusion and attention-based selection (see Fig. 2).

At first, each SKunit fuses volumes $\tilde{\mathbf{U}}^{(l)}$ and $\hat{\mathbf{U}}^{(l)}$ via element-wise summation, obtaining as a result $\mathbf{U}^{(l)} \in \mathbb{R}^{S\times S\times C}$ as $\mathbf{U}^{(l)} = \tilde{\mathbf{U}}^{(l)} + \hat{\mathbf{U}}^{(l)}$. Then, a global average pooling (GAP) is performed to generate the feature response vector (FRV) with the channel-wise statistics of the data, reducing the spatial dimension of $\mathbf{U}^{(l)}$ to $\mathbf{s}^{(l)} \in \mathbb{R}^C$ by taking the average of $S \times S$ spatial elements at each channel c:

$$s_c^{(l)} = \frac{1}{S^2}\sum_{i=1}^{S}\sum_{j=1}^{S} u_{i,j}^{(l),c} \tag{3}$$

The obtained FRV vector $\mathbf{s}$ is compacted by an FC layer defined by weights $\mathbf{W}_{fc}^{(l)} \in \mathbb{R}^{d\times C}$ followed by BN and ReLU, in order to obtain the neural activations of the different channel-features, enabling their guidance for adaptive kernel selections. In this sense, the feature weights vector (FWV) $\mathbf{z}^{(l)} \in \mathbb{R}^d$ can be defined as $\mathbf{z}^{(l)} = \text{ReLU}(\beta(\mathbf{W}_{fc}^{(l)} \cdot \mathbf{s}))$. Parameter d plays an important role in the performance of the SKunit, as its underestimation significantly reduces the efficiency of the MSKNet. For this reason, r is considered to control the compression rate of $\mathbf{z}^{(l)}$, being determined by $d = max(\frac{c}{r}, L)$, where $L = 32$ is the minimum value of d. Finally, to achieve an adaptive adjustment, a control gate is designed by means of an attention mechanism to select the most important regions of the FWV $\mathbf{z}^{(l)}$. This is done by applying one FC layer per SKunit's branch and computing the softmax function to obtain the effective FWV (EFWV) as:

$$a_c^{(l)} = \frac{e^{\mathbf{A}_c^{(l)}\cdot\mathbf{z}^{(l)}}}{e^{\mathbf{A}_c^{(l)}\cdot\mathbf{z}^{(l)}} + e^{\mathbf{B}_c^{(l)}\cdot\mathbf{z}^{(l)}}},\ b_c^{(l)} = \frac{e^{\mathbf{B}_c^{(l)}\cdot\mathbf{z}^{(l)}}}{e^{\mathbf{A}_c^{(l)}\cdot\mathbf{z}^{(l)}} + e^{\mathbf{B}_c^{(l)}\cdot\mathbf{z}^{(l)}}} \tag{4}$$

where $\mathbf{A}^{(l)}$, $\mathbf{B}^{(l)} \in \mathbb{R}^{C\times d}$ and $\mathbf{a}^{(l)}$, $\mathbf{b}^{(l)} \in \mathbb{R}^C$ are the soft attention vectors of $\tilde{\mathbf{U}}^{(l)}$ and $\hat{\mathbf{U}}^{(l)}$. Then, the final re-calibrated feature map $\mathbf{X}^{(l)}$ in the (l)-th SKunit is obtained by applying the attention-based vectors $\mathbf{a}^{(l)}$ and $\mathbf{b}^{(l)}$ along the channel dimension:

$$\mathbf{X}^{(l),c} = a_c^{(l)} \cdot \tilde{\mathbf{U}}^{(l)} + b_c^{(l)} \cdot \hat{\mathbf{U}}^{(l)},\ \text{subject to}\ a_c^{(l)} + b_c^{(l)} = 1 \tag{5}$$

Resulting $\mathbf{X}^{(l)}$ is fed to the next SKunit until the end of the network is reached, where a FC layer is applied to perform the final classification. Fig. 2 provides a detailed summary of the proposed model in terms of its layers, kernel size and output map dimensions. All weights are randomly initialized and trained using the back-propagation algorithm with Adam optimizer and cross-entropy loss. We use mini-batches of size 100, and train the network for 200 epochs without data augmentation.

III. Experimental Results

A. Hyperspectral Datasets

Three real HSI datasets have been considered in our experiments. The first one is the Indian Pines (IP) captured by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor [21] in 1992 over several agricultural fields in Northwestern Indiana. It comprises 16 different classes, with size of $145\times145\times200$. The second one is the University of Pavia (UP) scene, with size of $610\times340\times113$. This scene was gathered by the Reflective Optics Spectrographic Imaging System (ROSIS) sensor [22] over an urban area, comprising 9 different classes. The third one is the University of Houston (UH) scene [23], collected by the Compact Airborne Spectrographic Imager (CASI) in June 2012 over the University of Houston campus.

Its dimension is $349 \times 1905 \times 144$, containing 15 ground-truth classes. The UH scene was first presented by the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Technical Committee during the 2013 data fusion contest [24].

B. Experimental Setting

To evaluate the performance of the proposed MSKNet model for HSI classification, three widely used quantitative metrics have been considered: the overall accuracy (OA), average accuracy (AA), and Kappa coefficient. Moreover, the number of parameters has also been reported, to determine the volume of data to be trained and the computational cost. The proposed model has been compared with a recent study using different scale information in CNNs called MSDNet [25], and also with a standard CNN (Table I).

TABLE I
DETAILS OF STANDARD CNN FOR COMPARATIVE PURPOSES

Network	Layer ID	Kernel/Neurons	BatchNorm	Act. function
CNN	ConvA1	$bands \times 1 \times 1 \times 32$	Yes	Linear
	ConvB1	$32 \times 3 \times 3 \times 32$	Yes	ReLU
	ConvA2	$32 \times 1 \times 1 \times 64$	Yes	Linear
	ConvB2	$64 \times 3 \times 3 \times 64$	Yes	ReLU
	ConvA3	$64 \times 1 \times 1 \times 128$	Yes	Linear
	ConvB3	$128 \times 3 \times 3 \times 128$	Yes	ReLU
	FC	$n_{classes}$	No	Softmax

Different training percentages (i.e., varying the percentage of available labeled samples that are used for training, with the remaining labeled samples used for testing) and input patch size of 11×11 have been considered empirically. Fig. 3 shows the OA results obtained by the proposed method (denoted by MSKNet), MSDNet and the traditional CNN for the IP and UP datasets (in all cases, the vertical bar represent the average after 10 Monte Carlo experiments, with the error bar representing the standard variation). In general, the OA

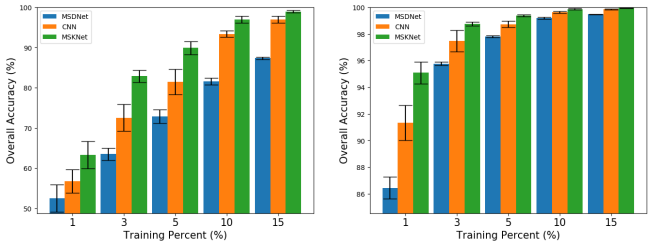

Fig. 3. OA results obtained by the proposed method, a standard CNN and MSDNet [25] for IP (a) and UP (b) scenes (average restults after 10 Monte Carlo experiments)

of the considered methods improves with the percentage of training samples. As it can be seen from Fig. 3(a), the proposed MSKNet outperforms the MSDNet and traditional CNN. The superiority of the proposed method over the other two methods is clear from Fig. 3, especially when the number of training samples is limited. In the IP dataset, the OA obtained for training percentages of 1%, 3% and 5% is poor due to the over-fitting problem. However, when the percentage increases, the proposed method achieves OA values higher than 95% (with smaller variance). Focusing on Fig. 3(b), the MSKNet method is superior to the traditional CNN method when the training data is limited. Specifically, the OA of MSKNet raises above 98% with 3% of training samples. Another interesting remark concerning Fig. 3(b) is that, when the training percentage is just 1%, the variance of the MSKNet is significantly lower than the other two methods. It is apparent that the performance of MSKNet consistently yields higher OAs than those obtained by the MSDNet and CNN.

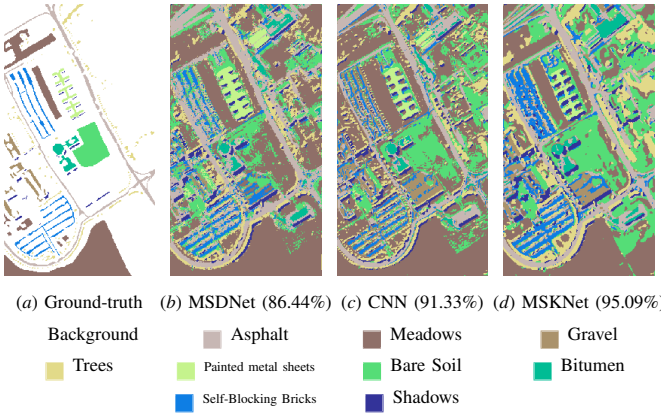

Fig. 4. Classification maps obtained for the UP scene (using 1% of the available labeled samples). The obtained OAs are shown in brackets.

Table II reports the classication accuracies of proposed method, MSDNet and traditional CNN for the considered datasets. Focusing in IP dataset, the OA, AA and Kappa of our MSKNet are higher than those reported by MSDNet and CNN, being MSDNet the worst classifier and CNN the thinnest method in terms of parameters. It is remarkable that the accuracies obtained for most classes by the proposed method are also higher than those obtained by the MSDNet and CNN. Regarding to UP dataset, the OA of the proposed method is increased to 90.66%, while the number of model parameters and the standard variation of the OA is the lowest. Most of the results obtained from two previous datasets hold true for UH dataset, as can be seen the OA, AA and kappa is increased to 88.28%, 88.87%, and 0.8728, respectively.

Fig. 4 shows the classication maps obtained by the proposed method and the traditional CNN on the UP image. As it can be observed, by incorporating the selective kernel unit, the visual quality of the proposed method is increased compared to the traditional CNN. For example, in Fig. 4(d), the structure of the building block is better preserved, with more clear boundaries as compared with the traditional CNN in Fig. 4(c). Also, by comparing the parking area in the bottom leftmost part of the image (area with trees and asphalt) in Figs. 4(c) and 4(d), we can observe that the regions obtained by our method are better connected.

IV. CONCLUSIONS AND FUTURE LINES

In this letter, we presented a new HSI classification framework that exploits different scales of information present in the input data. Our newly developed method, called MSKNet, creates different branches by convolving the input HSI data cubes with different kernel sizes, and then aggregates the resulting information using a non-linear attention mechanism. The classification results obtained by the proposed method on

TABLE II
COMPARISON BETWEEN THE STANDARD CNN AND THE PROPOSED METHOD USING THE FIXED TRAINING SET AVAILABLE FOR IP, UP AND UH SCENES IN HTTP://DASE.GRSS-IEEE.ORG.

Class	INDIAN PINES			UNIVERSITY OF PAVIA			UNIVERSITY OF HOUSTON		
	MSDNet	**CNN**	**MSKNet**	**MSDNet**	**CNN**	**MSKNet**	**MSDNet**	**CNN**	**MSKNet**
1	61.36±20.12	46.0±5.54	**65.33**±13.4	84.04±7.92	84.52±2.88	**84.96**±2.06	81.56±1.11	81.72±1.46	**82.75**±0.32
2	69.8±5.73	76.37±5.17	**85.98**±1.77	93.84±3.07	**96.3**±1.06	96.3±1.66	81.78±2.02	**86.62**±4.66	86.28±4.19
3	30.86±7.16	54.62±8.9	**68.85**±5.39	43.94±8.38	58.6±4.58	**64.32**±7.63	62.12±2.92	91.12±6.36	**95.25**±2.38
4	24.53±5.74	**45.62**±8.48	41.24±3.57	96.36±1.69	97.34±0.79	**97.42**±0.59	87.0±3.46	85.78±3.12	**90.29**±1.42
5	57.69±6.41	50.43±10.36	**69.77**±16.33	**99.03**±0.6	98.1±0.89	99.0±0.61	89.4±5.35	99.5±0.38	**99.8**±0.34
6	92.0±3.36	**93.69**±3.04	91.95±3.65	56.27±5.21	68.41±11.58	**77.42**±7.72	81.52±6.06	**90.33**±5.32	84.27±3.17
7	0.0±0.0	**0.0**±0.0	0.0±0.0	75.8±9.24	**87.15**±4.55	81.98±6.14	**82.78**±3.21	74.16±3.25	76.96±3.3
8	**96.44**±3.08	94.07±4.24	90.13±6.06	95.79±1.35	**96.14**±1.73	95.86±1.44	60.9±1.98	75.94±4.85	**77.92**±1.73
9	38.88±8.49	**80.0**±20.0	65.0±17.08	**96.62**±2.59	95.79±1.3	95.6±2.09	76.25±4.16	74.4±6.28	**84.48**±3.24
10	36.57±8.97	79.72±6.48	**85.55**±3.82	-	-	-	46.28±3.18	72.94±13.21	**84.38**±7.22
11	79.02±4.0	**86.98**±4.28	84.9±2.33	-	-	-	64.44±5.38	88.1±4.77	**94.21**±4.12
12	49.94±5.23	45.86±6.99	**54.43**±9.29	-	-	-	58.49±16.95	95.31±2.78	**98.27**±0.82
13	93.75±4.15	87.92±3.2	**93.96**±5.7	-	-	-	**88.35**±4.14	75.79±4.12	82.98±6.86
14	89.01±9.06	91.01±3.82	**95.57**±1.94	-	-	-	84.16±2.17	**99.6**±0.57	97.98±2.78
15	24.91±4.28	36.03±20.45	**65.66**±14.07	-	-	-	30.4±10.1	**98.34**±3.15	97.18±2.04
16	40.42±10.94	**89.77**±3.41	84.09±11.95	-	-	-	-	-	-
OA	66.49±1.28	76.5±1.92	**81.73**±1.92	85.84±1.56	89.43±1.61	**90.66**±1.32	71.57±2.95	84.56±1.4	**88.28**±1.31
AA	55.33±1.19	66.13±2.35	**71.4**±2.09	82.41±1.16	86.93±1.93	**88.09**±1.23	71.69±2.74	85.98±1.39	**88.87**±1.41
K(x100)	61.54±1.49	73.07±2.17	**79.2**±2.19	80.72±1.99	85.6±2.25	**87.34**±1.79	69.25±3.17	83.24±1.52	**87.28**±1.4
Parameters	1.54M	**263K**	322K	2.13M	237K	**202K**	2.57M	254K	**187K**

three real HSIs (with very limited training samples) outperform those achieved by the traditional CNN and the MSDNet quantitatively and also in terms of visual performance. In the future, we will combine our model with more sophisticated CNN architectures such as ResNet. We also plan to incorporate other different attention mechanisms to our model.

REFERENCES

[1] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, pp. S110 – S122, 2009, imaging Spectroscopy Special Issue.

[2] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–18, 2020.

[3] A. Signoroni, M. Savardi, A. Baronio, and S. Benini, "Deep learning meets hyperspectral image analysis: A multidisciplinary review," *Journal of Imaging*, vol. 5, no. 5, p. 52, May 2019.

[4] Y. Lee, H. Jung, D. Han, K. Kim, and J. Kim, "Learning receptive field size by learning filter size," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan 2019, pp. 1203–1212.

[5] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *IEEE Computer Vision and Pattern Recognition*, June 2019.

[6] M. P. Sceniak, D. L. Ringach, M. J. Hawken, and R. Shapley, "Contrast's effect on spatial summation by macaque v1 neurons," *Nature neuroscience*, vol. 2, no. 8, p. 733, 1999.

[7] R. Gaetano, D. Ienco, K. Ose, and R. Cresson, "A two-branch cnn architecture for land cover classification of pan and ms imagery," *Remote Sensing*, vol. 10, no. 11, p. 1746, Nov 2018.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[9] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectralspatial 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, Feb 2018.

[10] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. Plaza, J. Li, and F. Pla, "Capsule networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2145–2160, April 2019.

[11] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *31th AAAI Conference on Artificial Intelligence*, 2017.

[12] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5384–5394, Aug 2019.

[13] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[14] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectralspatial hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 740–754, Feb 2019.

[15] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," in *International Conference on Learning Representations*, 2017.

[16] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu, "Residual networks of residual networks: Multilevel residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1303–1314, June 2018.

[17] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 113, pp. 155–165, 2016.

[18] N. He, M. E. Paoletti, J. M. Haut, L. Fang, S. Li, A. Plaza, and J. Plaza, "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 755–769, Feb 2019.

[19] H. Gao, X. Zhu, S. Lin, and J. Dai, "Deformable kernels: Adapting effective receptive fields for object deformation," *arXiv preprint arXiv:1910.02940*, 2019.

[20] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 8065–8080, Oct 2019.

[21] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, M. R. Olah, and O. Williams, "Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," *Remote Sensing of Environment*, vol. 65, no. 3, pp. 227–248, 1998.

[22] B. Kunkel, F. Blechinger, R. Lutz, R. Doerffer, H. van der Piepen, and M. Schroder, "ROSIS (Reflective Optics System Imaging Spectrometer) - A candidate instrument for polar platform missions," in *Proc. SPIE 0868 Optoelectronic technologies for remote sensing from space*, J. Seeley and S. Bowyer, Eds., 1988, p. 8.

[23] X. Xu, f. Lil, and A. Plaza, "Fusion of hyperspectral and LiDAR data using morphological component analysis," in *2016 IEEE International Geoscience and Remote Sensing Symposium*, 2016, pp. 3575–3578.

[24] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pizurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2405–2418, June 2014.

[25] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9201–9222, Nov 2019.