Multi-Scale Depthwise Separable Convolutional Neural Network for Hyperspectral Image Classification

Jiliang Yan[1], Deming Zhai[1], Yi Niu[2], Xianming Liu[1], and Junjun Jiang[1](✉)

[1] School of Computer Science and Technology, Harbin Institute of Technology,
Harbin 150001, China
jiangjunjun@hit.edu.cn

[2] School of Artificial Intelligence, Xidian University, Xi'an 710071, China

Abstract. Hyperspectral images (HSIs) have far more spectral bands than conventional RGB images. The abundant spectral information provides very useful clues for the followup applications, such as classification and anomaly detection. How to extract discriminant features from HSIs is very important. In this work, we propose a novel spatial-spectral features extraction method for HSI classification by Multi-Scale Depthwise Separable Convolutional Neural Network (MDSCNN). This new model consists of a multi-scale atrous convolution module and two bottleneck residual units, which greatly increase the width and depth of the network. In addition, we use depthwise separable convolution instead of traditional 2D or 3D convolution to extract spatial and spectral features. Furthermore, considering classification accuracy can benifit from multi-scale information, we introduce atrous convolution with different dilation rates parallelly to extract more discriminant features of HSIs for classification. Experiments on three standard datasets show that the proposed MDSCNN has got the state-of-the-art accuracy among all compared methods.

Keywords: Hyperspectral images classification · Multi-scale · Depthwise separable convolution · Residual learning

1 Introduction

Recently, hyperspectral imaging technology has attracted widespread attention in the remote sensing society. The hyperspectral imager can capture accurate spectral response characteristics and spatial details of surface materials, which makes it possible to identify and classify the landcovers. HSI classification aims to assign a unique category to each pixel in the image, enabling automatic identification of categories and serving for following applications. However, due to the limit of labeled samples, the existence of mixed pixels, and the Houghes phenomenon, HSI classification is a very challenge problem.

G. Zhai et al. (Eds.): IFTC 2019, CCIS 1181, pp. 171–185, 2020.
https://doi.org/10.1007/978-981-15-3341-9_15

Based on traditional machine learning methods, many HSI classification approaches such as support vector machine (SVM) [1], multiple logistic regression [2], decision trees [3], *etc.* are proposed for pixel-level classification of HSI. However, HSIs usually provide hundreds of spectral bands, which contain a large amount of redundant information. Therefore, using raw spectral information directly not only results in high computational cost, but also reduces classification performance. Consequently, there are some methods that focus on mitigating the redundancy of HSIs with principal component analysis (PCA) [4] or linear discriminant analysis (LDA) [5]. Furthermore, spatial information has been reported to be very helpful in improving the representation of HSI data [6]. Thus more and more classification frameworks based on spatial-spectral features have been presented [7,8]. Although these spatial-spectral classification methods have achieved some progress, they all need to perform feature extraction engineering through human prior knowledge, which limits these methods in different scenarios.

Deep learning has become an important tool for big data analysis, and has made great breakthroughs in many computer vision tasks, such as image classification, object detection and natural language processing. Recently, it has been introduced into the HSI classification as a powerful feature extraction tool and shows great performance. Compared with the traditional artificial feature extraction methods, deep convolutional neural network can extract rich features from the original data through a series of layers. Since the learning process is completely automatic, deep learning is more suitable for dealing with complex scenes. Chen *et al.* [9] first applied the Stacked Autoencoder (SAE) to the HSI classification which is composed of multiple sparse autoencoders. Mughees *et al.* [10] proposed a Spectral-Adaptive Segmentation DBN (SAS-DBS) for HSI classification that exploits the spatial-spectral features by segmenting the original spectral bands into small sets and processing each group separately by local DBNs. However, deep neural networks such as SAE are based on the fully connected layer. Although the above-mentioned deep neural network models can effectively extract deep features in HSIs, they may ignore the spatial information of HSIs. Unlike SAE, Convolutional Neural Networks (CNN) can directly extract spatial and spectral features of HSIs while keeping the input shape. For this reason, most of the current HSI classification networks with spatial-spectral features are based on CNN structure. They can be divided into two main categories. The first is to extract spatial and spectral features separately and then combine them and feed to the classifier [11]. Another strategy is to extract the spatial-spectral joint features of HSIs simultaneously by 3D convolution [12,13]. Although these methods can effectively extract the spatial spectral information of HSIs, they all ignore the multi-scale characteristics. Because of the complexity and diversity of HSI scenery, it is often difficult to extract spatial information from a single scale.

In this work, we propose a Multi-Scale Depthwise Separable Convolutional Neural Network (MDSCNN) for HSI classification which can effectively exploit spatial-spectral features and achieve competitive HSI classification performance.

This new model consists of a multi-scale atrous convolution module and two bottleneck residual units, which greatly increase the width and depth of the network. In addition, we use depthwise separable convolution instead of traditional 2D or 3D convolution, which leads to extract spectral features poorly or has high computational complexity. In contrast, the depthwise separable convolution can not only extract spatial-spectral features separately, but also greatly reduce the amount of training parameters. Furthermore, considering classification accuracy can benifit from multi-scale information, we introduce atrous convolution with different dilation rates parallelly to extract more discriminant features of HSIs. Experiments on three standard datasets show that the proposed MDSCNN has got the state-of-the-art accuracy among all compared methods.

The remainder of this paper is organized as follows: In Sect. 2 several related techniques are described. Section 2 introduces the proposed MDSCNN model. The experiments and results analysis are shown in Sect. 3, A conclusion is made in Sect. 4.

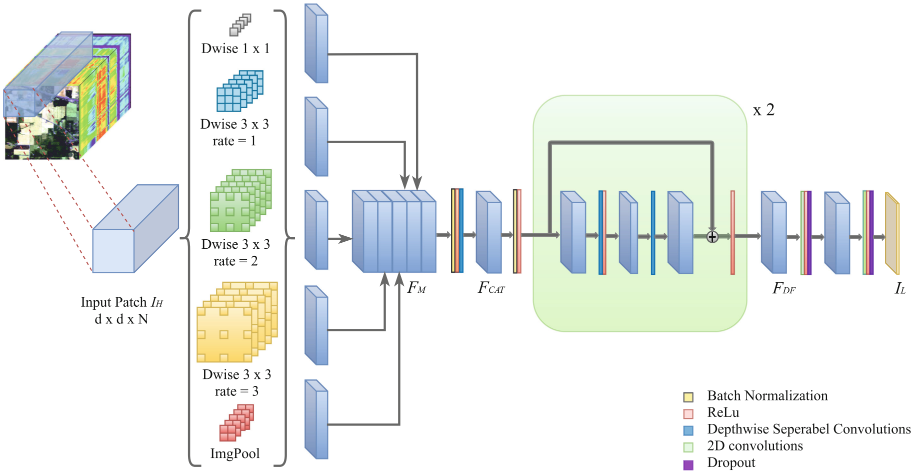

Fig. 1. Overview of the proposed Multi-Scale Depthwise Separable CNN (MDSCNN) model.

2 Method

In this section, we have discussed the overall architecture of the proposed MDSCNN firstly. Then we provide a detailed explanation about each module.

2.1 Overall Architecture

We have constructed a wide and deep network with a specially developed multi-scale atrous convolution module and two depthwise separable bottleneck residual units for HSI classification. As shown in Fig. 1, the proposed MDSCNN is a fully convolutional network (FCN) [19] without any fully connected layers, so that it can handle any input patches with arbitrary size and produce the same size output. Let's denote I_H and I_L as the input HSI patch and predicted labels of MDSCNN.

Table 1. The proposed MDSCNN topology. M is the number of bands.

Module	Layer	Input channel	Output channel	Kernel size	Padding	Parameters
Multi-scale atrous conv module	DW 1×1	M	M	1×1	0	$1 \times 1 \times M$
	PW 1×1	M	128	1×1	0	$1 \times 1 \times M \times 128$
	DW 3×3 (r = 1)	M	M	3×3	1	$3 \times 3 \times M$
	PW 1×1	M	128	1×1	0	$1 \times 1 \times M \times 128$
	DW 3×3 (r = 2)	M	M	3×3	2	$3 \times 3 \times M$
	PW 1×1	M	128	1×1	0	$1 \times 1 \times M \times 128$
	DW 3×3 (r = 3)	M	M	3×3	3	$3 \times 3 \times M$
	PW 1×1	M	128	1×1	0	$1 \times 1 \times M \times 128$
	ImgPool	M	M	2×2	0	0
		M	128	1×1	0	1×1×M×128
Concatenate Conv	Conv	640	128	1×1	0	$1 \times 1 \times 640 \times 128$
Depthwise separable bottleneck residual unit	Layer1 DW 3×3	128	128	3×3	1	$3 \times 3 \times 128$
	Layer1 PW 1×1	128	64	1×1	0	$1 \times 1 \times 128 \times 64$
	Layer2 DW 3×3	64	64	3×3	1	$3 \times 3 \times 64$
	Layer2 PW1 $\times$ 1	64	64	1×1	0	$1 \times 1 \times 64 \times 64$
	Layer3 DW 3×3	64	64	3×3	1	$3 \times 3 \times 64$
	Layer3 PW 1×1	64	128	1×1	0	$1 \times 1 \times 64 \times 128$
Classification module	Conv 1	128	128	3×3	1	$3 \times 3 \times 128 \times 128$
	Conv 2	128	128	3×3	1	$3 \times 3 \times 128 \times 128$
	Conv 3	128	C_Number	1×1	0	$1 \times 1 \times 128 \times C\_Number$

The input to the spectral pixel based methods usually is a pixel vector $x^{1 \times 1 \times M}$, where M is the number of spectral bands. In order to simultaneously exploit the spatial and spectral features, it is necessary to introduce a three-dimensional approach to incorporate the contextual information. In this method, we feed the network with a $d \times d$ patch P centered on x, where d is the width and height of the patch. In this way, the original spatial and spectral features can be considered simultaneously. Especially, the model is designed to predict the label of center pixel, whose position index is $[d/2+1, d/2+1, M]$. Meanwhile, we need to select the value of d carefully. It will result in lacking spatial information if d is too small. On the other hand, when we set d too large, it may introduce some pixels that are not belong to the same class. Furthermore, a multi-scale atrous convolution module is introduced to extract rich spatial and spectral features. It extracts multi-scale features F_M from I_H

$$F_M = H_{MAC}(I_H), \tag{1}$$

where $H_{MAC}(\cdot)$ denotes multi-scale atrous convolution operation. F_M is a joint maps with multi-scale features, then F_M is concatenated together via one 1×1 Conv layer

$$F_{CAT} = W_{CAT}(F_M), \tag{2}$$

where $W_{CAT}(\cdot)$ and F_{CAT} denote the weight set to the Conv layer and joint features respectively. The following are backbone of the network, two specially designed Depthwise Separable bottleneck Residual (DSR) units, implemented with depthwise separable convolution

$$F_{DF} = H_{DSR}(H_{DSR}(F_{CAT})), \tag{3}$$

where $H_{DSR}(\cdot)$ denotes our residual unit, F_{DF} is the obtained deep discriminative feature. The end of the model are three convolutional layers for classification, and we insert the dropout layer (p = 0.5) during training to prevent overfitting

$$I_L = H_{CLS}(F_{DF}), \tag{4}$$

where $H_{CLS}(\cdot)$ and I_L denate classification module and label map predicted by MDSCNN.

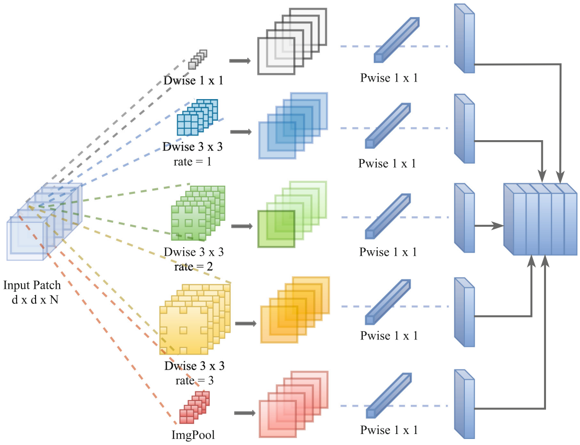

Fig. 2. The developed multi-scale atrous convolution module.

In this paper, we select the cross-entropy as the loss function to train the network, which can be formulated as:

$$E = -\sum_{i=1}^{T} y_i \cdot \log h\left(x_i\right), \tag{5}$$

where T denotes the total number of training samples, y_i is the ground truth of x_i, and $h(\cdot)$ denotes the softmax function which is computed as:

$$h_{x_i} = \frac{e_i^z}{\sum_{j=1}^{C} e_j^z}, \tag{6}$$

where z_i is the features learned from sample x_i, and C is the number of label categories.

We have summarized the proposed MDSCNN in Table 1, which includes the number of channels, kernel size, padding value and parameters for each convolution or pooling layer of each module.

2.2 Multi-scale Feature Extraction with Atrous Convolution

It has been proved that classification can benifit from abundant contextual features [20]. Inspired by the Atrous Spatial Pyramid Pooling (ASPP) module [15] which is commonly used in semantic segmentation, we design a multi-scale atrous convolution module based on depthwise separable convolution. As shown in Fig. 2, it consists of four filters: 1×1, 3×3 $(r=1)$, 3×3 $(r=2)$, 3×3 $(r=3)$, and an $ImagePooling$ branch. Atrous convolution can enlarge the receptive field of the filter while maintaining the amount of parameters. These convolutions are extracted in parallel with different dilation rates, and then the generated feature maps are concatenated together. Therefore, we pad the input patches to ensure the shape of generated feature maps are same. Early studies have shown that a 3×3 atrous convolution with an extremely large rate will degenerate into a simple 1×1 convolution. In this way, it will not be able to capture long range information due to image boundary effects [15]. Therefore, considering the spatial size d of input patch is generally between 9–25, we set the maximum dilation

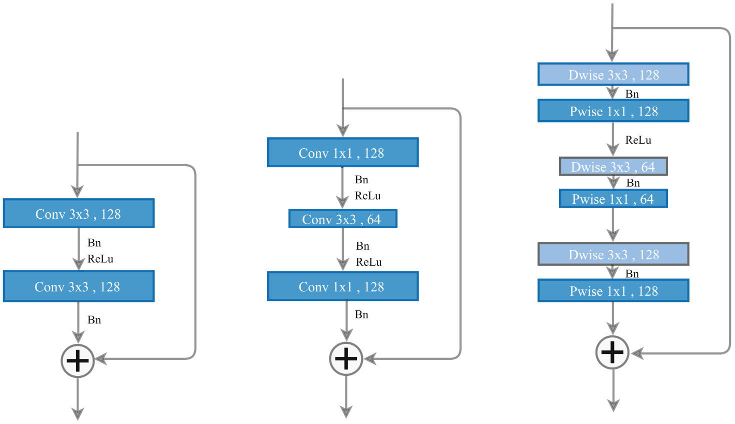

Fig. 3. Different residual unit architectures. (Left) Traditional residual units, (Middle) Bottleneck residual units, (Right) The developed Depthwise Separable bottleneck Residual (DSR) units.

rate to 3. In addition, The pooling filter preserves the image-level features of the original HSIs and enriches the diversity of features.

2.3 BottleNeck Residual Block with Depthwise Separable Convolution

Deep convolutional neural networks often appear degradation phenomenon due to inadequate training, therefore, He *et al.* [21] constructed an identity mapping to ease the training process. The basic idea is that if E is a perfect network with best performance, the T is a deeper network with some redundant layers, so the goal is to make redundant layer become an identical transformation. That is to say, T's performance is the same as E. Therefore, the network needs to learn a residual $F(x) = H(x) - x$, where x is original feature, $H(x)$ denotes the features learned from x. $H(x)$ will be equivalent with x if the network learns nothing, *i.e.* $F(x) = 0$. Since fitting the residual $F(x)$ is easier than fitting the original $H(x)$, residual network can effectively avoid degradation of network. The design of the residual units becomes a point worth exploring, as we can see, there are three different residual units showed in Fig. 3. Basic residual unit (Left) contains two convolutional layers. Bottleneck residual unit (Middle) [22] is more economical than the conventional residual block, and its input and output feature maps dimension is first reduced and then restored, which reduces the calculation amount of the middle layer and allows a faster execution.

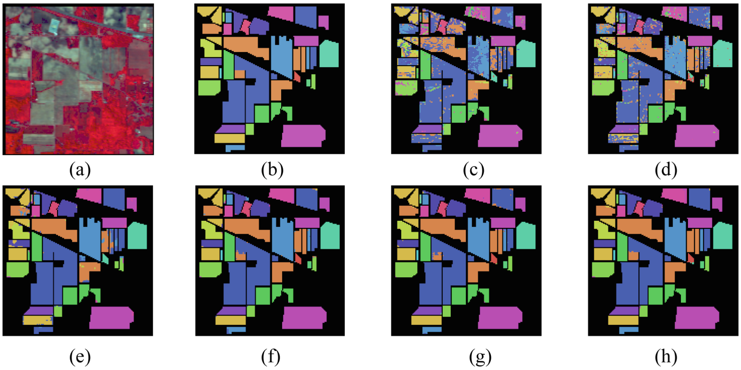
(a) (b) (c) (d)
(e) (f) (g) (h)

Fig. 4. Classification maps for IP dataset. (a) Simulated RGB composition of the scene. (b) Ground-Truth classification map. Classification maps obtained by (c) MLP, (d) SVM, (e) 2D-CNN, (f) 3D-CNN, (g) HybridSN, (h) MDSCNN

As we know, these traditional residual units mainly focus on the spatial features of RGB images, the spectral features are not well extracted. Inspired by bottleneck, here we have specially designed a DSR unit for HSI classification as shown in Fig. 3 (Right). It mainly consists of three convolutional layers, which extract spatial features using depthwise convolution firstly, and then convolute point by point to extract spectral features. After the first convolutional layer, the feature map dimension is reduced (from the number of channels point of view). A nonlinear activation function is introduced between the first and second convolutional layers

$$F_{MID} = W_{pw} W_{dw} \sigma(W_{pw} W_{dw} \cdot x), \tag{7}$$

where W_{pw} and W_{dw} denate the weights of pointwise convolution and depthwise convolution respectively. $\sigma(\cdot)$ denotes the ReLu activation function. F_{MID} is the feature map after the second Conv layer in DSR. Since the depthwise convolution's output is shallow, in order to retain as much information as possible, a linear output is put between the second and third convolutional layers without adding any nonlinear activation function. Thus a output F_{DF} is obtained via the shortcut connection:

$$F_{DF} = W_{pw} W_{dw} \cdot F_{MID} + x, \tag{8}$$

where + is an elementwise addition that does not change the size of the feature map.

The experimental results show that adding two depthwise separable convolutional units improves the classification accuracy while using limited training samples.

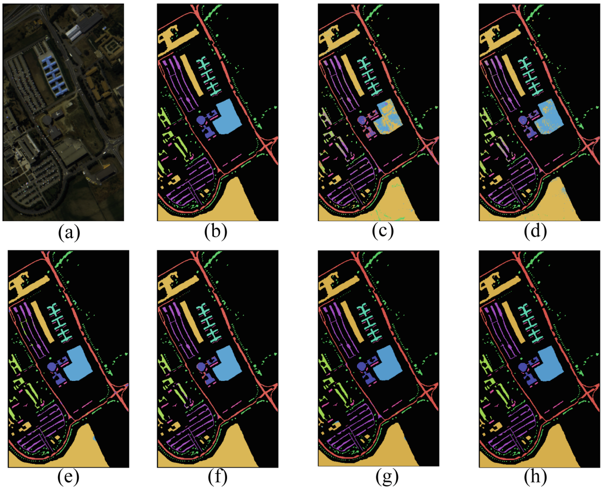

Fig. 5. Classification maps for UP dataset. (a) Simulated RGB composition of the scene. (b) Ground-Truth classification map. Classification maps obtained by (c) MLP, (d) SVM, (e) 2D-CNN, (f) 3D-CNN, (g) HybridSN, (h) MDSCNN

3 Experiments

3.1 Experimental Datasets

We have evaluated our model on three well-know HSI datasets, which are widely used for HSI classification, Indian Pines (IP), University of Pavia (UP) and Salinas Valley (SV).

IP: This scene was gathered by AVIRIS sensor in North-western Indiana, which consists of 145 × 145 pixels and 224 spectral reflectance bands in the wavelength range from 400 nm to 2500 nm. We have also reserved the number of bands to 200 by removing 24 damaged bands.

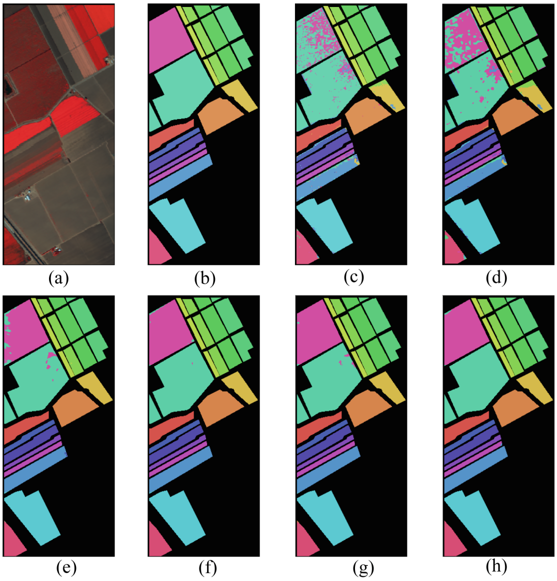

Fig. 6. Classification maps for UP dataset. (a) Simulated RGB composition of the scene. (b) Ground-Truth classification map. Classification maps obtained by (c) MLP, (d) SVM, (e) 2D-CNN, (f) 3D-CNN, (g) HybridSN, (h) MDSCNN

UP: This dataset captured the urban area around the University of Pavia, Italy. The spatial resolution of the image is 1.3 m per pixel, the spectral coverage ranges from 0.43 m to 0.86 m. After 12 bands is removed due to noise, there are 103 bands left. The image consists of 610 × 340 pixels, but it contains many background pixels.

Table 2. Classification results for IP dataset using 15% of the labeled data for training and 15 × 15 input spatial size.

Class	Training/Test	SVM	MLP	2D-CNN	3D-CNN	HybirdSN	MDSCNN
Alfalfa	6/46	72.50	63.63	0.00	75.86	95.65	**95.65**
Corn-notill	214/1428	81.24	72.15	92.45	93.10	98.47	**99.72**
Corn-mintill	124/830	81.14	77.02	**99.21**	94.65	98.67	97.87
Corn	35/237	73.45	64.71	73.96	87.83	99.58	**100.00**
Grass-pasture	72/483	90.56	81.42	93.36	97.38	98.74	**100.00**
Grass-trees	109730	93.08	91.44	99.17	98.38	99.05	**100.00**
Grass-pasture-mowed	4/28	100.00	100.00	0.00	99.04	89.28	**100.00**
Hay-windrowed	71/478	95.36	89.33	90.87	**99.58**	99.17	98.97
Oats	3/20	70.00	100.00	100.00	78.26	95.24	**100.00**
Soybean-nottill	145/972	78.63	69.91	94.92	97.64	99.37	**99.49**
Soybean-mintill	368/2455	79.52	70.04	92.95	98.61	99.02	**99.79**
Soybean-clean	88/593	80.42	58.74	90.18	97.22	98.17	**99.32**
Wheats	30/205	94.74	89.03	97.15	98.55	98.56	**100.00**
Woods	189/1265	93.36	91.35	93.71	95.01	**99.92**	98.21
Building-Grass-Trees	57/386	79.61	73.74	93.80	95.97	96.98	**98.92**
Stone-Steel-Towers	13/93	97.78	98.90	98.93	97.85	98.94	**98.94**
OA		81.96	76.62	96.27	97.84	98.88	**99.33**
AA		85.09	80.71	81.92	94.12	97.80	**99.45**
Kappa		81.96	73.11	92.52	95.75	98.72	**99.23**

SV: The SV dataset was captured by an onboard visible/infrared imaging spectrometer over Salinas Valley,California. The image has 512 × 217 pixels with a spatial resolution of 3.7 m per pixel. The image originally contained 224 bands, but the remaining 204 bands were usually used for experiments after removing 20 water absorption bands.

Table 3. Classification results for UP dataset using 15% of the labeled data for training and 15 × 15 input spatial size.

Class	Training/Test	SVM	MLP	2D-CNN	3D-CNN	HybirdSN	MDSCNN
Asphalt	994/6631	94.32	88.72	96.61	99.14	**99.89**	99.77
Meadows	2797/18649	95.28	86.02	98.03	98.45	**99.97**	98.47
Gravel	314/2099	84.91	78.88	**100.00**	90.19	99.43	99.81
Trees	459/3064	97.95	87.39	98.49	98.89	99.19	**99.71**
Sheets	201/1345	99.48	99.48	100.00	100.00	100.00	**100.00**
Bare soils	754/5029	93.40	91.90	100.00	100.00	98.28	**100.00**
Bitumen	199/1330	89.65	84.70	99.32	86.24	100.00	**100.00**
Bricks	552/3682	85.19	76.27	94.16	99.53	95.00	**99.95**
Shadows	142/947	100.00	100.00	100.00	99.79	99.58	**100.00**
OA		93.82	86.38	97.98	98.14	98.52	**99.26**
AA		93.35	88.15	98.51	96.92	99.04	**99.75**
Kappa		91.76	81.48	97.34	97.56	98.04	**99.03**

In this paper, we implement the proposed method with Pytorch framework. Before training, we have enhanced the data by randomly flipping and adding noise. We use the Adam optimizer to train the network with a batch size of 64 and initially set a base learning rate as 0.001 then reduce it with poly (0.9).

3.2 Experimental Results

We compare the proposed MDSCNN with several classical and state-of-the-art HSI classification methods. (1) SVM; (2) MLP; (3) 2D-CNN [23]; (4) 3D-CNN [12]; (5) HybidSN [13]. SVM and MLP both are spectral-based methods. 2D-CNN is based on spatial features. 3D-CNN utilizes spatial-spectral features of HSIs with 3D convolution, which consists of two 3D convolutional layers and one fully connected layer. The HybridSN firstly get a low-dimensional data with PCA as input, and it contains three 3D convolutional layers, one 2D convolutional layer, and three fully connected layers in the end of model. We evaluate all of these methods on three standard datasets described above. In order to evaluate the proposed MDSCNN and demonstrate the effectiveness of the multi-scale strategy, we have conducted the following three experiments.

(1) In our first experiment: the first step is to randomly divide the original IP,UP and SV dataset respectively into two subsets: the training set and testing set, whose sample numbers are shown in the first column of Tables 2, 3 and 4. We train all methods mentioned above with some optimal parameters. In addition, for our model, the input patch size is set to $15 \times 15 \times M$.
(2) In our second experiment: intuitively, different spatial size of patch has significant effect on the classification performance of model. We takes three different sizes of patch as input: $9 \times 9, 15 \times 15, 21 \times 21$ and 15% of the available training data for experiment.
(3) In our third experiment: to verify the effectiveness of the multi-scale atrous convolution module used to jointly extract the mutli-scale spatial-spectral features, we compare the proposed MDSCNN to the network without the multi-scale module. To verify the effectiveness of the DSR units, we also compare the performance of the proposed MDSCNN to a similar network with the DSR unit replaced with traditional two convolutional layers residual unit.

To evaluate the performance of different methods, three objective metrics: overall accuracy (OA), average accuracy (AA), and the Kappa coefficient, are adopted in these experiments.

Experiment 1: Tables 2, 3 and 4 show the quantitative results, moreover, the best result is highlighted in bold font. As shown, the results of traditional spectral-based pixel-level classification methods are not satisfactory, like SVM and MLP, which are far worse than spatial-spectral based methods like 2D-CNN, 3D-CNN, HybridSN. The proposed MDSCNN achieves the best classification accuracy on each dataset. Furthermore, the proposed MDSCNN has improved

Table 4. Classification results for SV dataset using 15% of the labeled data for training and 15 × 15 input spatial size.

Class	Training/Test	SVM	MLP	2D-CNN	3D-CNN	HybirdSN	MDSCNN
Brocoli-green-weeds-1	302/2009	100.00	100.00	99.87	100.00	100.00	**100.00**
Brocoli-green-weeds-2	559/3726	99.71	98.59	99.97	99.84	99.97	**100.00**
Fallow	297/1976	98.64	95.40	99.83	99.92	**99.92**	99.85
Fallow-rough-plow	210/1394	98.58	98.86	99.21	99.78	**99.85**	99.50
Fallow-smooth	402/2678	98.92	91.49	99.22	**99.78**	99.18	99.51
Stubble	594/3959	99.98	99.98	99.94	99.92	99.95	**100.00**
Celery	537/3579	99.78	99.02	99.66	**100.00**	99.80	99.92
Graphes-untrained	1691/11271	79.21	68.92	90.04	98.86	94.68	**99.34**
Soil-vinyard-develop	931/6203	99.18	97.14	99.69	99.96	99.66	**100.00**
Corn-senesced-green-weeds	492/3278	95.46	87.82	99.33	**99.96**	99.69	99.91
Lettuce-romaine-4wk	161/1068	97.71	90.77	94.19	**100.00**	99.98	99.91
Lettuce-romaine-5wk	290/1927	98.32	97.27	99.43	99.84	99.94	**100.00**
Lettuce-romaine-6wk	138/916	99.12	94.23	99.89	99.93	100.00	**100.00**
Lettuce-romaine-7wk	161/1070	98.03	95.64	97.43	99.90	100.00	**100.00**
Vinyard-untrined	1091/7268	83.69	80.02	78.22	82.95	**99.98**	99.01
Vinyard-vertical-trellis	272/1807	99.89	97.51	99.83	**100.00**	99.93	99.63
OA		92.65	87.69	94.45	97.18	98.70	**99.87**
AA		96.64	93.29	97.53	98.86	99.55	**99.78**
Kappa		91.80	86.21	95.14	96.86	98.57	**99.86**

the OA value 0.5% at least compared to the suboptimal method on all testing sets, and there are surprising improvements in AA and Kappa. In addition to the qualitative results, the Figs. 4, 5 and 6 show three visual classification maps of the different methods on three datasets respectively. It can be observed that the traditional single-pixel-based methods have a lot of noise due to the lack of spatial information. Meanwhile, the classification maps obtained by the spatial-spectral based methods are smoother, and the most of the ewrong classified pixels exist around the boundaries of some categories. Taking all these observations into account, it is possible to state that the MDSCNN provides a more accurate and robust classification result than all of the other tested methods.

Table 5. Results on the proposed MDSCNN when considering different spatial size input patches.

Spatial size	IP			UP			SV		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
9 × 9	98.41	97.11	98.19	98.79	99.15	98.40	98.57	99.27	98.42
11 × 11	98.91	99.08	98.75	99.23	99.94	98.98	98.73	99.57	98.58
15 × 15	**99.33**	**99.45**	**99.23**	**99.27**	**99.96**	**99.03**	**99.87**	**99.93**	**99.86**
21 × 21	98.62	97.71	98.43	99.24	99.95	98.99	97.17	98.86	96.86

Experiment 2: Table 5 shows the classification results of the proposed MDSCNN when using different spatial size patches as input. The OA, AA, and Kappa values all increase firstly and then decrease on the three datasets. The highest score was reached as the spatial size is set to 15 × 15. It is not difficult to understand that increasing the spatial size of the patch will introduce a certain amount of spatial information at first. But as the spatial size continues to increase, a lot of noise or pixels with different classes will be also introduced.

Table 6. Classification performance comparison of the proposed MDSCNN and the network without multi-scale atrous convolution module and the network with traditional two layers residual unit.

Method	UP			IP		
	OA	AA	Kappa	OA	AA	Kappa
w/o MS	94.71	93.80	93.02	89.65	88.23	88.92
w/o DRS	96.34	96.05	95.29	92.51	92.64	92.36
MDSCNN	**99.26**	**99.75**	**99.03**	**99.33**	**99.45**	**99.23**

Experiment 3: As shown in Table 6, the multi-scale atrous convolution module outperforms the network without it (by 4.55% for the UP dataset, 9.68% for the IP dataset in OA classification performance). Beyond that, our developed DRS units achieve better performance than traditional residual units.

4 Conclusion

In this paper, a novel multi-scale separable convolutional network for HSI classification is proposed, the model leverages a multi-scale atrous convolutional module to extract spatial-spectral features from a HSI patch. In addition, a specially designed depthwise separable bottleneck residual unit is applyed to increase the depth of the network and improve classification performance. The proposed MDSCNN is deep while it doesn't introduce large quantities of training parameters because of the depthwise separable convolution. The final experimental results show that our method achieves outstanding classification performance with a relative small number of training samples. Although multi-scale features fusion has been adopted in our MDSCNN, the features at different stages of the network is not considerated. In the future, we will continue to explore some new multi-stage information fusion ways for HSI classification.

Acknowledgements. This work is supported by the National Science Foundation under Grant Nos. 61922027, 61672193, and 61971165.

References

1. Mercier, G., Lennon, M.: Support vector machines for hyperspectral image classification with spectral-based kernels. In: 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477), IGARSS 2003, vol. 1, pp. 288–290, July 2003
2. Li, J., Bioucas-Dias, J.M., Plaza, A.: Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. IEEE Trans. Geosci. Remote Sens. **48**(11), 4085–4098 (2010)
3. Li, S., Zhang, B., Gao, L., Zhang, L.: Classification of coastal zone based on decision tree and PPI. In: 2009 IEEE International Geoscience and Remote Sensing Symposium, vol. 4, pp. IV-188–IV-191, July 2009
4. Chen, H., Chen, C.H.: Hyperspectral image data unsupervised classification using Gauss-Markov random fields and PCA principle. In: IEEE International Geoscience and Remote Sensing Symposium, vol. 3, pp. 1431–1433, June 2002
5. Bandos, T.V., Bruzzone, L., Camps-Valls, G.: Classification of hyperspectral images with regularized linear discriminant analysis. IEEE Trans. Geosci. Remote Sens. **47**(3), 862–873 (2009)
6. Demir, B., Ertürk, S.: Improving SVM classification accuracy using a hierarchical approach for hyperspectral images. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 2849–2852, November 2009
7. Fauvel, M., Tarabalka, Y., Benediktsson, J.A., Chanussot, J., Tilton, J.C.: Advances in spectral-spatial classification of hyperspectral images. Proc. IEEE **101**(3), 652–675 (2013)
8. Wang, J., Jiao, L., Wang, S., Hou, B., Liu, F.: Adaptive nonlocal spatial-spectral kernel for hyperspectral imagery classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **9**(9), 4086–4101 (2016)
9. Chen, Y., Lin, Z., Zhao, X., Wang, G., Gu, Y.: Deep learning-based classification of hyperspectral data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **7**(6), 2094–2107 (2014)
10. Mughees, A., Tao, L.: Multiple deep-belief-network-based spectral-spatial classification of hyperspectral images. Tsinghua Sci. Technol. **24**(2), 183–194 (2019)
11. Yang, G., Gewali, U.B., Ientilucci, E., Gartley, M., Monteiro, S.T.: Dual-channel DenseNet for hyperspectral image classification. In: 2018 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018, pp. 2595–2598, July 2018
12. Hamida, A.B., Benoit, A., Lambert, P., Amar, C.B.: 3-D deep learning approach for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **56**(8), 4420–4434 (2018)
13. Roy, S.K., Krishna, G., Dubey, S.R., Chaudhuri, B.B.: HybridSN: exploring 3D–2D CNN feature hierarchy for hyperspectral image classification. ArXiv, abs/1902.06701 (2019)
14. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. CoRR, abs/1406.4729 (2014)
15. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. CoRR, abs/1606.00915 (2016)
16. LeCun, Y., et al.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on

Neural Information Processing Systems, NIPS 2012, vol. 1, pp. 1097–1105. Curran Associates Inc., USA (2012)

18. Szegedy, C., et al.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, June 2015
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CoRR, abs/1411.4038 (2014)
20. Huang, G., Chen, D., Li, T., Wu, F., van der Maaten, L., Weinberger, K.Q.: Multi-scale dense convolutional networks for efficient prediction. CoRR, abs/1703.09844 (2017)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR, abs/1512.03385 (2015)
22. Tishby, N., Zaslavsky, N.: Deep learning and the information bottleneck principle. CoRR, abs/1503.02406 (2015)
23. Liu, B., Yu, X., Zhang, P., Tan, X., Yu, A., Xue, Z.: A semi-supervised convolutional neural network for hyperspectral image classification (2017)