Deformable Convolutional Neural Networks for Hyperspectral Image Classification

Jian Zhu, Leyuan Fang, *Senior Member, IEEE*, and Pedram Ghamisi, *Member, IEEE*

***Abstract*— Convolutional neural networks (CNNs) have recently been demonstrated to be a powerful tool for hyperspectral image (HSI) classification, since they adopt deep convolutional layers whose kernels can effectively extract high-level spatial–spectral features. However, sampling locations of traditional convolutional kernels are fixed and cannot be changed according to complex spatial structures in HSIs. In addition, the typical pooling layers (e.g., average or maximum operations) in CNNs are also fixed and cannot be learned for feature downsampling in an adaptive manner. In this letter, a novel deformable CNN-based HSI classification method is proposed, which is called deformable HSI classification networks (DHCNet). The proposed network, DHCNet, introduces the deformable convolutional sampling locations, whose size and shape can be adaptively adjusted according to HSIs' complex spatial contexts. Specifically, to create the deformable sampling locations, 2-D offsets are first calculated for each pixel of input images. The sampling locations of each pixel with calculated offsets can cover the locations of other neighboring pixels with similar characteristics. With the deformable sampling locations, deformable feature images are then created by compressing neighboring similar structural information of each pixel into fixed grids. Therefore, applying the regular convolutions on the deformable feature images can reflect complex structures more effectively. Moreover, instead of adopting the pooling layers, the strided convolution is further introduced on the feature images, which can be learned for feature downsampling according to spatial contexts. Experimental results on two real HSI data sets demonstrate that DHCNet can obtain better classification performance than can several well-known classification methods.**

***Index Terms*— Convolutional neural networks (CNNs), deformable convolution, hyperspectral image (HSI) classification, spatial–spectral feature extraction.**

I. Introduction

HYPERSPECTRAL images (HSIs) consist of hundreds of spectral bands, which span from the visible to infrared spectrum. Each pixel in an HSI is a high-dimensional vector, and its entries correspond to the spectral responses of various spectral bands. Such rich spectral information can be used for effective classification in a wide variety of applications, such as agriculture, military surveillance, and environment monitoring [1], [2].

Manuscript received January 9, 2018; revised March 18, 2018; accepted April 19, 2018. This work was supported in part by the National Natural Science Foundation under Grant 61771192 and Grant 61471167, and in part by the National Natural Science Foundation for Young Scientist of China under Grant 61501180. *(Corresponding author: Leyuan Fang.)*

J. Zhu and L. Fang are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: jianzhu@hnu.edu.cn; fangleyuan@gmail.com).

P. Ghamisi is with the German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF), 82234 Wessling, Germany, and also with the Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: p.ghamisi@gmail.com).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2018.2830403

Over the last few decades, numerous types of pixelwise-based classification models (e.g., support vector machine (SVM) [3], multinomial logistic regression [4], and active learning [5]) have been developed. In general, these methods [3]–[5] take full advantage of spectral information of HSIs to obtain good classification results. However, the classification maps are still noisy since the spatial contexts are not considered. Later, many spatial–spectral classification models [6]–[12] were proposed to improve the classification performance. For instance, superpixel [6], [7] and multiple kernel learning [8] were developed to improve the support vector machine classifier based on spatial–spectral information. In [6], a multiscale superpixel segmentation was used to model the distribution of classes based on the spatial information for an SVM classifier. In [8], three kernels were separately employed for the utilization of the spatial–spectral information, and then they were combined together for classification. Sparse representation [9], [10] also demonstrates to be a powerful tool for HSI classification. The spatial–spectral information of HSIs in a neighboring region can be incorporated into a sparse model.

Recently, convolutional neural networks (CNNs) have made great breakthrough in many fields, such as object detection image classification [13] and natural language processing [14]. Attempts [15]–[17] have been made to apply CNNs on HSIs to extract features and achieve outstanding performance. In [15], a framework based on off-the-shelf CNNs was proposed for remote sensing image scene classification. In [16], a deep feature extraction architectures based on a CNN with kernels sampling on input images are proposed to extract spectral–spatial features of HSI. In general, the sampling locations of the convolutional kernels mentioned above are of a fixed grid and the pooling layer cannot be trained to learn feature downsampling. However, the HSIs have very complex structures of different scales or shapes. Therefore, the traditional CNN-based methods may not effectively extract the features from the complex structures in HSIs, which limits the classification performance.

To address the above issues, in this letter, a novel deformable CNN (DCNN) [18]-based HSI classification method is proposed, which is called as deformable HSI classification networks (DHCNet). Different from the traditional HSI classification model, the DHCNet introduces the deformable convolutional sampling locations. To create the deformable sampling locations, 2-D offsets are first calculated for each pixel of the input image. The sampling locations of each pixel with calculated offsets can cover the locations of other neighboring similar pixels. With the deformable sampling locations, deformable features can be obtained by compressing the neighboring similar structural information of each pixel into fixed grids. The deformable feature images can appropriately reflect complex structures, and as a result, the regular convolutions applied on them can also reflect

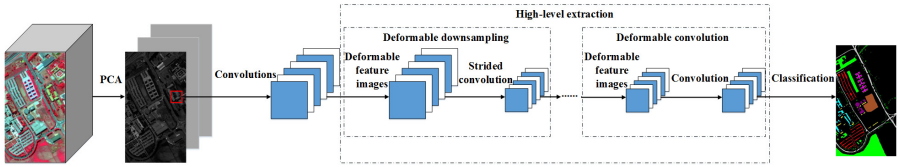

Fig. 1. Illustration of the architecture of the DHCNet.

different spatial structures in an adaptive manner. In addition, due to the fact that the traditional pooling layer cannot be learned, the strided convolution with deformable sampling locations is further introduced on the feature images, which can be learned for downsampling according to spatial contexts.

The remaining of this letter is organized as follows. Section II reviews the CNN-based HSI classification model and introduces the proposed DHCNet method. Experimental results and analysis on two real HSIs are provided in Section III. Finally, Section V concludes this letter.

II. Proposed DHCNet Method

A. CNN-Based HSI Classification Model

A CNN classifier [15] adopts a stack of layers to extract the features of an input image and assign a specific class label to each particular pixel. The layers commonly consist of a number of convolution layers, pooling layers, and fully connected layers.

In CNNs, the inputs are image patches (of size $N \times N \times C$) which are centered at the available labeled pixels. These samples are utilized to train CNNs, since the neighboring pixels may have similar spatial contexts. In the CNNs, K_1 different filter kernels (of size $n \times n$) spatially slide over the input samples in the first convolutional layers. In this manner, a 3-D volume of feature maps (of size $M \times M \times K_1$) can be obtained. After passing through multiple convolutions, the number of feature maps increases, which lead to high computational burdens for latter layers. Therefore, pooling layers conduct feature downsampling operations (max or averaging) after the convolutional layers to reduce the dimensions of the feature maps. After a series of convolutional and pooling layers, the fully connected layers combine the output values of all in the previous layers into an n-dimensional vector to extract high-level features. Finally, the outputs of the last fully connected layer, whose output nodes are equal to the number of classes, are fed to the softmax layer to generate the probability distribution that the pixel belongs to each class.

The CNNs can automatically learn high-level features through a stack of layers and demonstrate to be powerful for HSI classification. However, the convolutional kernels and pooling operators of CNNs cannot be adaptively adjusted with respect to spatial features, since the sampling locations of the convolutions are fixed grids and pooling layers cannot be learned for feature downsampling.

B. DHCNet

To solve the issues above, we propose a novel DHCNet model in this letter. We introduce the deformable convolutional sampling locations in high-level feature extraction, which was first adopted in [18] in the computer vision community and they replaced the last several convolutional layers with deformable convolutional layers. With deformable sampling locations, deformable feature images can be obtained, which fuse the neighboring similar structural information of each pixel in an adaptive manner. On the one hand, by applying regular convolutions on deformable feature maps, one can extract the complex spatial context in a more effective way. On the other hand, strided convolutions are adopted on them to be learned for feature downsampling according to the structural information instead of traditional pooling layers. The flowchart and the architecture of the proposed DHCNet are illustrated in Fig. 1 and Table I, respectively.

TABLE I
Architecture of Models

Model	Layers	Type	Filter number
CNNs DCNNs DHCNet	1-2	Convolutional	96
	3	Pooling	-
	4-5	Convolutional	108
	6	Pooling	-
	7-8	Convolutional	128
	9	Global Pooling	-
	10	Fully connected	256
	11	Dropout	-
	12	Fully connected	Class number

First, the principal component analysis is performed on original HSIs to extract the first three informative components. Then, HSIs are segmented into patches to be considered as the training sample for DHCNet. Note that the label of the central pixel in a patch represents the label of the patch. To more effectively extract the complex spatial–structural information, we adopt the deformable convolutions and deformable downsampling in the high-level extraction process, which are described as follows.

1) Deformable Convolution: The sampling locations of regular convolution are fixed grids over a 2-D image (of size $N \times N$), denoted by

$$s = (x, y)|(0, 0), (0, 1), \ldots, (N-1, N-1).$$

However, as illustrated in Fig. 3(a), the regular convolutional filters cannot appropriately cover the feature structure with various shapes (e.g., circle and rhombus). To address this issue, we introduce the deformable sampling location for convolutional filters. Two steps are added before regular convolution, including: 1) creation of offset field and 2) creation of deformable feature maps. As shown in Fig. 2, the offset fields [18] are first calculated for each pixel by convolutions over the input feature images, whose channels are twice as many as the input feature images. Here, given a pixel in an input feature image with location (x, y) and value $p(x, y)$, it corresponds to two values Δx and Δy in the offset fields. Then, a deformable feature image is generated to fusing the information of neighboring similar pixels. The pixel value is

$$p_{\text{new}}(x, y) = p(x_n, y_n) \tag{1}$$

where $x_n = \min(\max(0, x + \Delta x), N-1)$, $y_n = \min(\max(0, y+\Delta y), N-1)$. x_n and y_n are fractional locations, and the value of $p(x_n, y_n)$ is calculated according to values of four surrounding integer locations via bilinear interpolation. The weights of the convolutional filters for generating offsets fields are trained based on spatial features to enable

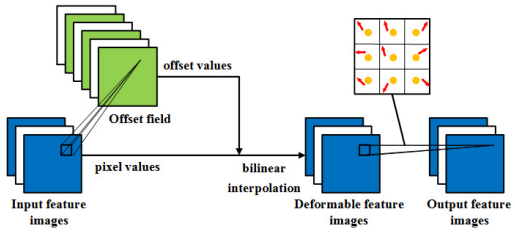

Fig. 2. Illustration of the deformable convolution. The offset field is obtained by applying regular convolutions on the input feature images to create deformable feature maps. Therefore, the sampling locations of standard convolutional filters can be transferred to the neighboring similar pixels.

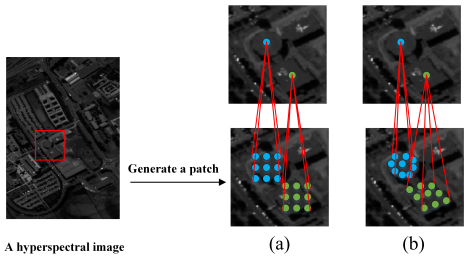

Fig. 3. Illustration of the difference between the sampling locations of (a) regular convolution and (b) deformable convolution.

the sampling locations to be transferred to the neighboring similar pixels. Finally, regular convolutions are operated on the deformable feature images and the output can be obtained as follows:

$$y(x, y) = \Sigma_{s_w} w_{ij} \cdot p_{\text{new}(x_{ij}, y_{ij})} \tag{2}$$

where s_w enumerates all the locations of the kernel and w_{ij} represents the corresponding weight of the kernel. In Fig. 3(b), sampling locations of deformable convolutional filters are present in the shape of circle and rhombus, better covering the HSIs' feature structures.

2) Deformable Downsampling: Downsampling can reduce the size of input feature images to accelerate the learning process and obtain high-level features. However, the traditional pooling layers cannot be learned, and sampling locations of them are also a fixed grid, which cannot effectively extract the structural information. Instead of adopting typical pooling layers, applying strided convolution on deformable feature images mentioned earlier can be learned for feature downsampling. Moreover, with deformable sampling locations, the kernel weights of the downsampling can be trained with more effectively spatial structure information.

Compared with images in computer vision, the training patches of HSIs have much smaller sizes with limited detail. In addition, the model is more likely to suffer from overfitting with much fewer training samples available for HSIs than what we have in computer vision. Therefore, we adopt the deformable convolution and deformable downsampling in the high-level feature extraction of HSIs. Here, pixels fuse the information of many pixels in original images, where the feature extraction has a great effect on the classification accuracy. Moreover, a limited number of parameters added to the model will create not much computational cost for the training process. To further accelerate the training procedure, the batch normalization is adopted after each convolutional layer, which also can stabilize the training process of the offsets of the deformable convolutional layers.

After utilizing a series of convolution layers, a global pooling layer is used to reduce the parameters of the model.

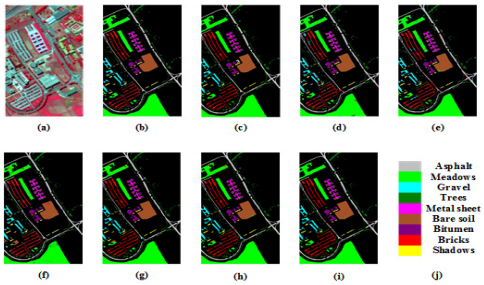

Fig. 4. Reference map and classification result of the University of Pavia image. (a) False-color composite image. (b) Reference. (c) SVM-IID. (d) EPF. (e) SC-MK. (f) MNFL. (g) CNNs. (h) DCNNs. (i) DHCNet. (j) Color code.

Then, a fully connected layer is adopted to learn the feature extracted by the convolutional layers, and a dropout layer is used after the fully connected layer to reduce the overfitting. Finally, another fully connected layer determines the class of the pixel belonging to. For the output feature vector z, the probability distribution can be denoted by

$$p_i = \frac{e^{z_i}}{\Sigma_{j=1}^{k} e^{z_j}}, \quad i = 1, 2, \ldots, k \tag{3}$$

where k represents the number of the class. The label of a test hyperspectral pixel t can be predicted by the maximal probability

$$\text{Class}(t) = \arg \max_{i=1,2,\ldots,k} p_i. \tag{4}$$

III. Experimental Results

A. Data Set Description

The proposed DHCNet[1] is evaluated on two real HSI data sets: University of Pavia and Houston University. The size of the University of Pavia image is $610 \times 340 \times 103$ with a spatial resolution of 1.3 m/pixel. The spectral bands of the image range from 430 to 860 nm while 12 very noisy bands are removed. The data set has nine classes of interest [11]. Fig. 4(a) and (b) shows the false-color composite of the University of Pavia image and the corresponding reference data. The Houston University data set contains 144 spectral bands ranging from 380 to 1050 nm of size 349×1905, whose spatial resolution is 2.5 m/pixel. The data set consists of 15 classes of interest [11]. Fig. 5(a) and (b) illustrates the false-color composite of the Houston University image and the corresponding reference data.

B. Compared Classifiers

The performance of the proposed framework is compared with that of several well-known HSI classification methods: SVM-intrinsic image decomposition (SVM-IID) [21], edge-preserving filter (EPF) [19], superpixel-based classification via multiple kernels (SC-MKs) [8], multiple nonlinear feature learning (MNFL) with multivariate logistic regression [20], original CNNs, and DCNNs. The parameters of the SVM-IID, EPF, SC-MK, and MNFL are set to the default values of their original works [21], [19], [8], [20].

Note that for a fair comparison, the original CNNs, the DCNNs, and the proposed DHCNet have the similar architecture, which is shown in Table I. The original CNN is a plain network whose extraction layers consist of regular convolutional layers and maximum pooling layers. The DCNNs replace convolutional layers in high-level extraction

[1]Source code is released at: https://github.com/OrdianryCore/DHCNet

TABLE II

AAs (IN %) OF TEN EXPERIMENTS ON THE UNIVERSITY OF PAVIA IMAGE WITH DIFFERENT TRAINING SAMPLES OBTAINED BY EACH METHOD. THE STANDARD DEVIATION VALUES ARE GIVEN IN BRACKETS

	Tr=45			Tr=55			Tr=65		
Method	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
SVM-IID	96.54(1.01)	96.95(0.70)	95.56(1.35)	97.08(0.91)	97.23(0.29)	96.38(1.22)	97.86(1.44)	97.88(0.60)	97.32(0.18)
EPF	92.00(2.23)	91.22(2.09)	89.61(2.80)	94.18(1.14)	92.95(1.62)	92.35(1.46)	94.69(1.89)	93.49(1.97)	93.03(2.43)
SC-MK	95.53(0.75)	96.73(0.49)	94.10(0.97)	95.95(0.78)	97.17(0.52)	94.65(1.02)	96.84(0.52)	97.62(0.26)	95.81(0.67)
MNFL	94.89(0.98)	96.12(0.58)	93.28(1.25)	95.42(0.89)	96.77(0.40)	93.99(1.14)	96.02(0.98)	97.13(0.54)	94.76(1.27)
CNNs	96.53(0.90)	96.82(0.57)	95.42(1.17)	97.09(0.96)	97.36(0.57)	96.16(1.24)	97.43(0.80)	97.80(0.39)	96.60(1.05)
DCNNs	96.97(0.84)	97.30(0.51)	96.00(1.10)	97.65(0.63)	97.83(0.50)	96.89(0.83)	98.17(0.49)	98.29(0.35)	97.58(0.65)
DHCNet	**97.37**(0.60)	**97.70**(0.44)	**96.52**(0.78)	**97.95**(0.75)	**98.03**(0.50)	**97.28**(0.99)	**98.44**(0.46)	**98.54**(0.26)	**97.93**(0.61)

TABLE III

AAs (IN %) OF TEN EXPERIMENTS ON THE HOUSTON UNIVERSITY IMAGE WITH DIFFERENT TRAINING SAMPLES OBTAINED BY EACH METHOD. THE STANDARD DEVIATION VALUES ARE GIVEN IN BRACKETS

	Tr=30			Tr=40			Tr=50		
Method	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
SVM-IID	89.97(1.01)	90.18(0.94)	88.82(0.97)	91.66(0.89)	92.01(0.93)	91.13(0.74)	93.08(0.49)	93.16(0.50)	92.50(0.45)
EPF	90.85(1.26)	90.55(1.42)	90.10(1.36)	92.01(2.01)	91.80(1.99)	91.36(2.17)	93.97(0.87)	93.93(1.05)	93.48(0.94)
SC-MK	90.01(1.14)	91.11(0.87)	89.20(1.23)	92.52(1.03)	93.24(0.78)	91.92(1.11)	93.21(0.86)	93.69(0.67)	92.66(0.93)
MNFL	88.55(1.21)	89.42(0.92)	87.62(1.31)	91.00(1.34)	91.65(0.98)	90.27(1.45)	92.82(0.64)	93.30(0.57)	92.23(0.69)
CNNs	92.25(1.14)	93.50(1.02)	91.62(1.23)	94.22(0.60)	95.15(0.52)	93.75(0.65)	95.50(0.42)	96.36(0.41)	95.32(0.45)
DCNNs	92.87(0.97)	94.02(0.99)	92.29(1.05)	94.93(0.84)	95.76(0.73)	94.51(0.91)	96.11(0.36)	96.77(0.35)	95.79(0.39)
DHCNet	**93.50**(0.98)	**94.52**(0.93)	**92.97**(1.06)	**95.21**(0.61)	**95.98**(0.58)	**94.82**(0.66)	**96.37**(0.26)	**97.00**(0.25)	**96.08**(0.28)

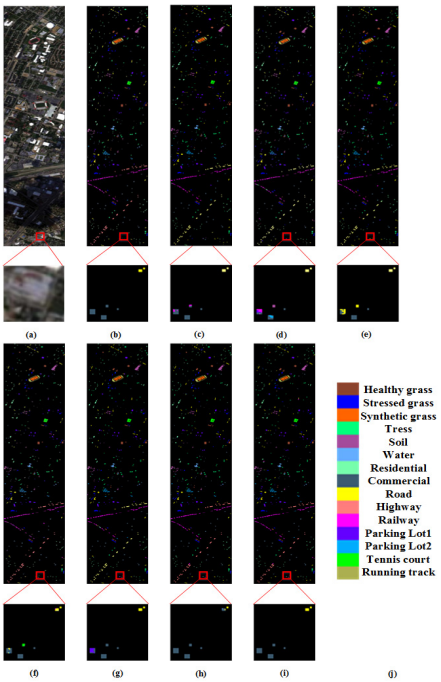

Fig. 5. Reference map and classification result of the Houston University image. (a) False-color composite image. (b) Reference. (c) SVM-IID. (d) EPF. (e) SC-MK. (f) MNFL. (g) CNNs. (h) DCNNs. (i) DHCNet. (j) Color code.

with the deformable convolutional layers. DHCNet further adopts the deformable strided convolutional layer instead of pooling layers in the last downsampling layer. For the above three networks, the initial learning rate is set as 0.1 with a momentum 0.9 and learning rate decay 0.25 every 500 steps. The size of training minibatch is 150, and the three networks are trained for 1500 iterations.

In order to quantitatively evaluate the classification performance, three objective metrics are adopted including overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa) [8].

The first experiment is performed on the University of Pavia image. The training set consists of 45, 55, and 65 samples, respectively, randomly selected per class. The remaining samples are used as the testing set. Fig. 4 illustrates the classification maps (Tr = 65) obtained by different approaches. As can be observed, the proposed DHCNet can more accurately classify pixels in the near edge regions and provide very similar results to the reference map compared with other methods. Table II shows the classification accuracies of the DHCNet and the other studied methods. The classification accuracies are the average value on ten experiments with random training samples. As can be seen, deep learning-based methods (e.g., CNNs, DCNNs, and DHCNet) generally obtain better results than do other traditional methods. In addition, the DCNNs achieve better classification results than original CNNs, since DCNNs adopt the deformable convolution layer. Moreover, the DHCNet leads to an improvement using the deformable strided convolutional layer and performs better than other compared methods in terms of OA, AA, and K.

The second experiment is conducted on the Houston University image with different numbers of training samples selected from each class (i.e., 30, 40, and 50 samples, respectively). The classification maps obtained by DHCNet and other compared methods with Tr = 50 are shown in Fig. 5, and the classification accuracies of each method are reported in Table III. As can be observed, DHCNet also outperforms other methods and achieves the OA improvement of 1.25%, 0.99%, and 0.87%, respectively, compared with original CNNs.

The assessment of the significance of the classification results obtained by the CNNs, the DCNNs, and the DHCNet based on McNemar's test [22] is given in Table IV for the University of Pavia (Tr = 65) and Houston University (Tr = 50). As can be observed, the classification results obtained by the DCNNs are statistically significant compared with the CNNs. The results of the proposed DHCNet are also statistically significant compared with the CNNs and the DCNNs.

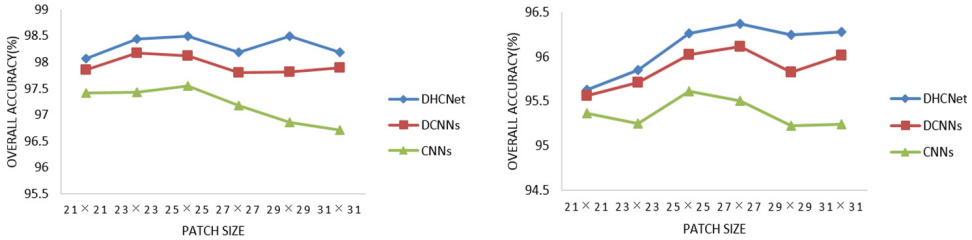

Fig. 6. Effect of the patch size on classification accuracies in the case of (a) University of Pavia image and (b) Houston University image.

TABLE IV

ASSESSMENT OF THE SIGNIFICANCE OF CLASSIFICATION ACCURACIES (Z) OBTAINED BY THE CNNS, THE DCNNS, AND THE DHCNET ON THE UNIVERSITY OF PAVIA AND HOUSTON UNIVERSITY IMAGES

	University of Pavia		Houston University	
	CNNs	DCNNs	CNNs	DCNNS
DCNNs	14.59	-	4.87	-
DHCNet	17.67	7.26	6.89	3.15

C. Effects of Patch Size on Classification Accuracies

The CNNs, the DCNNs and the proposed DHCNet have been evaluated on two images with Tr = 65 and 50 for University of Pavia and Houston University, respectively. As shown in Fig. 6(a) and (b), the patch size fromx 21 × 21 to 25 × 25, the CNNs will demonstrate better performance. As the patch size further increases, the performance of the CNNs will deteriorate. The performance of the DCNN and the proposed DHCNet methods will first improve, when the patch size increases to 29 × 29 and 27 × 27 on the University of Pavia and Houston University, respectively. In general, the DCNNs and the proposed DHCNet can outperform the CNNs on all the patch sizes. In addition, as the size of the patch enlarges, the improvement of the DCNN and proposed DHCNet methods over the CNNs is more obvious. The main reason is that the patches with bigger size contain more detailed structure information, and therefore DHCNet can effectively extract the spatial context useful for the classification step.

IV. CONCLUSION

In this letter, a novel DCNN framework was proposed to exploit spatial information for HSI classification, which is called DHCNet. The DHCNet method introduces the deformable sampling locations. On the one hand, with deformable sampling locations, deformable feature images were created and regular convolutions have been applied on them to extract features. On the other hand, strided convolutions are further adopted on the feature images to be learned for feature downsampling. In this way, more effective spatial features can be utilized in high-level extraction according to the complex structure information. The experimental results demonstrate that DHCNet performs better than several well-known HSI classification in terms of both the quality of the classification map and classification accuracy.

REFERENCES

[1] B. Luo, C. Yang, J. Chanussot, and L. Zhang, "Crop yield estimation based on unsupervised linear unmixing of multidate hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 162–173, Jan. 2013.

[2] X. Yang and Y. Yu, "Estimating soil salinity under various moisture conditions: An experimental study," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2525–2533, May 2017.

[3] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[4] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[5] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Trans.Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3947–3960, Oct. 2011.

[6] H. Yu, L. Gao, W. Liao, B. Zhang, A. Pižurica, and W. Philips, "Multiscale superpixel-level subspace-based support vector machines for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2142–2146, Nov. 2017.

[7] L. Fang, H. Zhuo, and S. Li, "Super-resolution of hyperspectral image via superpixel-based sparse representation," *Neurocomputing*, vol. 273, pp. 171–177, Jan. 2018.

[8] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson, "Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6663–6674, Dec. 2015.

[9] J. Zou, W. Li, and Q. Du, "Sparse representation-based nearest neighbor classifiers for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2418–2422, Dec. 2015.

[10] L. Fang, C. Wang, S. Li, and J. A. Benediktsson, "Hyperspectral image classification via multiple-feature-based adaptive sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1646–1657, Jul. 2017.

[11] L. Fang, N. He, S. Li, P. Ghamisi, and J. A. Benediktsson, "Extinction profiles fusion for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1803–1815, Mar. 2018.

[12] L. Fang, N. He, S. Li, A. J. Plaza, and J. Plaza, "A new spatial-spectral feature extraction method for hyperspectral images using local covariance matrix representation," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2018.2801387.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[14] Y. Wei *et al.*, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Dec. 2016.

[15] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.

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

[17] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[18] J. Dai *et al.* (2017). "Deformable convolutional networks." [Online]. Available: https://arxiv.org/abs/1703.06211

[19] X. Kang, S. Li, and J. A. Benediktsson, "Spectral–spatial hyperspectral image classification with edge-preserving filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2666–2677, May 2014.

[20] J. Li *et al.*, "Multiple feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1592–1606, Mar. 2015.

[21] X. Kang, S. Li, L. Fang, and J. A. Benediktsson, "Intrinsic image decomposition for feature extraction of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2241–2253, Apr. 2015.

[22] B. Rasti, P. Ghamisi, and R. Gloaguen, "Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3997–4007, Jul. 2017.