Multiscale CNNs Ensemble Based Self-Learning for Hyperspectral Image Classification

Leyuan Fang, *Senior Member, IEEE*, Wenke Zhao, Nanjun He, *Student Member, IEEE*, and Jian Zhu

***Abstract*— Fully supervised methods for hyperspectral image (HSI) classification usually require a considerable number of training samples to obtain high classification accuracy. However, it is time-consuming and difficult to collect the training samples. Under this context, semisupervised learning, which can effectively augment the number of training samples and extract the underlying information among the unlabeled samples, gained much attention. In this letter, we propose a Multiscale convolutional neural networks (CNNs) Ensemble Based Self-Learning (MCE-SL) method for semisupervised HSI classification. Generally, the proposed MCE-SL method consists of the following two stages. In the first stage, the spatial information of different scales from limited labeled training samples are extracted to train several CNN models. In the second stage, the trained multiscale CNNs are used to classify the unlabeled samples. After error correction, the problem of label partially incorrect is alleviated, and unlabeled samples with high confidence will be added to the original training data set for the next training iteration. We conduct comprehensive experiments on two real HSI data sets, and the experimental results show that the proposed MCE-SL can obtain better classification performance compared with several traditional semisupervised methods in few iterations.**

***Index Terms*— Convolutional neural network (CNN), ensemble approach, hyperspectral image (HSI) classification, self-learning, semisupervised learning (SSL).**

I. Introduction

DIFFERENT from the natural image that only contains three color channels, hyperspectral image (HSI) consists of hundreds of feature bands, which contains rich information in both spatial and spectral dimension, and can detect more latent details. Therefore, HSIs are widely used in disaster detective, land planning, and the environment monitoring [1]. Generally, these applications always need accurate classification of HSI. To achieve high accuracy of HSI classification, many fully supervised classification methods are proposed. Support vector machines (SVMs) [2] and random forest [3] are two typical supervised techniques to solve the problem of HSI classification. Deep neural networks based methods [4]–[6] become the mainstream in recent years since it can classify the HSI by an end-to-end way, and achieve a higher accuracy. However, fully supervised classification methods always require a considerable number of training samples to obtain high classification accuracy, which is time-consuming, and difficult to collect the training samples for HSI since it needs to label the samples on the pixel-level [7].

To address this problem, researchers try to augment the data set based on a few labeled samples [8]. Haut *et al.* [9] adopt random occlusion to generate different samples for increasing the variations of spatial features. Specifically, active learning (AL) is proposed to utilize the prior knowledge to learn new information and gather experience, in turn, knowledge and experience interact constantly [10]. AL attempts to select the most useful unlabeled samples through a particular query strategy to obtain a better classifier. Generally, the query strategy is based on uncertainty and diversity [11], such as Maximum Entropy [12], Mutual Information Criterion [8], and so on. Those functions are illustrated and contrasted by Haut *et al.* [13].

Compared with the AL algorithm that is an interaction process between the system and experts, self-learning algorithm does not rely on the human's intervention. Through self-learning, the classifier trained by the labeled samples will be utilized to assign a unique label to the unannotated data that belongs to a candidate set, where it can help to improve the performance of model gradually. Dopido *et al.* [14] transform AL to a self-learning framework, in which the trained model can automatically select the most valuable unlabeled samples, and label them. In the process of self-learning, many works fuse the spatial and spectral information to enhance the classification performance. In [15] and [16], the clustering analysis is used to combine spatial and spectral information. In [17] and [18], the random walk is utilized to make use of the spatial information for post processing. In addition to the traditional methods, the emerging neural network algorithm shows a very promising classification performance. Wu and Prasad [19] proposed a constrained Dirichlet process mixture model for semisupervised clustering, and treat the spectral band as sequence so that the convolutional recurrent

Manuscript received March 27, 2019; revised September 21, 2019; accepted October 23, 2019. This work was supported in part by the National Natural Science Foundation under Grant 61922029 and Grant 61771192, in part by the National Natural Science Fund of China for International Cooperation and Exchanges under Grant 61520106001, and in part by the Fund of Hunan Province for Science and Technology Plan Project under Grant 2017RS3024. *(Corresponding author: Nanjun He.)*

Leyuan Fang, Wenke Zhao, and Nanjun He are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China, and also with the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Changsha 410082, China (e-mail: fangleyuan@gmail.com; douhe@hnu.edu.cn; henanjun@hnu.edu.cn).

Jian Zhu is with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China, and also with the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai 7201804, China (e-mail: jianzhu@tongji.edu.cn).

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.2019.2950441

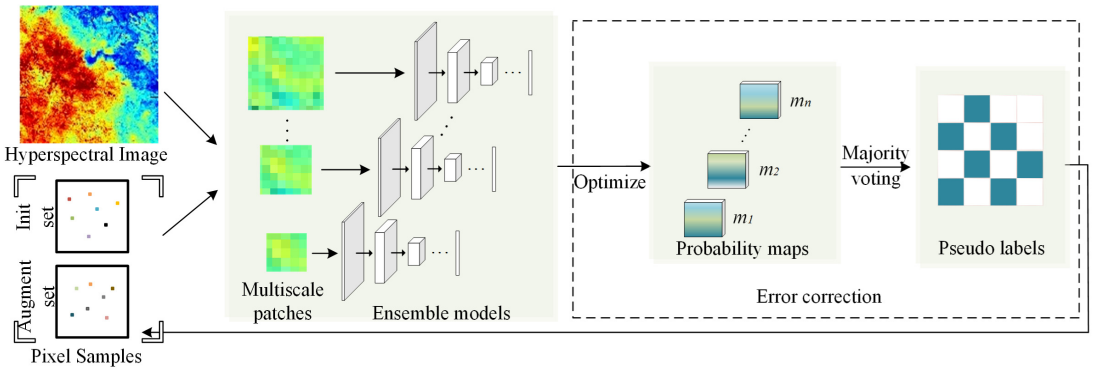

Fig. 1. Framework of the proposed select strategy. Multiscale patches with the same centering pixel sample are used as the input of ensemble models to get the probability maps. An error correct mechanism is utilized to alleviate the label partially incorrect problem and select the pseudo samples.

neural network (CRNN) can be used. However, the above-mentioned methods only explore the spatial information in one single fixed scale, which may not fully exploit the spatial information of HSI.

In this letter, we propose a multiscale convolutional neural networks (CNNs) ensemble based self-learning (MCE-SL) method for semisupervised HSI classification. Specifically, a few labeled training samples with spatial information of different scales are used to train several CNN models, in which different classifiers can complement each other in performance. In the second step, the multiscale classifiers are utilized to assign a label to each unlabeled sample, and the extend random walk (ERW) algorithm is applied to optimize the results. Finally, a majority voting based strategy is conducted to select the unlabeled samples with high confidence, which can alleviate the label partially as incorrect problem. Those selected instances will be added to original training data set for the next training iteration.

The remainder of this letter is organized as follows. Section II introduces the proposed method. Section III presents and analyzes the experimental results. In Section IV, we conclude this letter.

II. Proposed Method

The proposed method mainly consists of two steps: training of ensemble models, and collection of unlabeled samples. The first stage aims to obtain several classifiers that learned high-level features, which have the capacity to judge the candidate samples set. Then, an error correction mechanism based on majority voting is used to select the most informative and confident samples to augment the initial training set. Self-learning is an iteration process until the final result is satisfied. The framework of the proposed method is shown in Fig. 1.

A. Multiscale CNNs Ensemble

The amount of training samples and its distribution in space are both important to improve the classification accuracy. In this letter, we propose a new method based on multiscale spatial information and majority voting, which selects the unlabeled samples in an effective way and enhances its confidence.

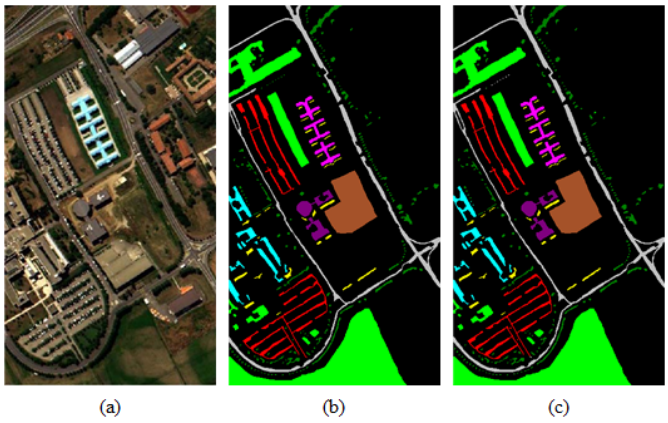

Fig. 2. University of Pavia. (a) Color image. (b) Ground truth. (c) Classification map of labeled samples.

The high dimensional image has rich spectral information, while it also has very high computational complexity. Therefore, the principal component analysis (PCA) [20], [21] is first conducted to project the initial image into m-bands. The PCA algorithm can remove the redundant information and preserve the less relevant information, i.e., the most useful information [22]. This operation can greatly reduce the parameters in CNN, and the dimension reduced image will be used in the construction of weight graph in ERW in the follow-up step.

Next, n different rough CNN classifiers are trained by a few labeled samples with n different patch sizes, respectively. Specifically, we labeled the patch of each sample with the label of central pixel in the patch [22]. It is based on the assumption that the adjacent pixels often have the same label so that the local spatial information and spectral information are imported to the model at the same time. In addition, patches with different sizes will show different prediction accuracy for a certain instances since their spatial information are discrepant. Therefore, n submodels are trained with multiscale patches, where patches belonging to the same sample have the same center pixel point. Each classifier will independently distinguish all class and the outputs are presented as a manner of probability maps. All those models will be integrated as an entirety, they exploit the complementary information among different scales.

B. Error Correction Mechanism

The initial probability maps estimated by the above process still have noises for the rough CNN models and self-learning's automatic annotation. Meanwhile, the patch only utilizes the local spatial information, and the spatial relationship between different patches is not used. In order to alleviate the label partially incorrect problem, ERW algorithm is used to optimize the initial probability maps. Thus, n optimized probability maps $M = (m_1, m_2, \ldots, m_n)$ with the size of (h, w, c) are obtained, where h, w, c indicate the height, width, and the number of classes of the HSI image, respectively.

Let the $P_{ij}(0 \leq i \leq n, j \in D_u)$ be the optimized probability vector that a pixel j belongs to each class, where D_u represents the unlabeled data set, and n is the number of optimized probability maps. Therefore, we can calculate the mean probability vector P_{mj} of each pixel and the new mean probability map M_m, in which

$$P_{mj} = \frac{1}{N} \sum_{i=0}^{n} P_{ij}. \tag{1}$$

The set of M and M_m form a committee, and every member give its own judgment to the unlabeled samples. Specifically, the class label $L_{ij} = argmaxP_{ij}$ to each pixel in the M and the voting label $L_m = argmaxP_{mj}$ to each pixel in the M_m are obtained, respectively. The pseudo-labels are selected based on the majority voting, that is, the samples accepted by most of the committee member will be considered. To further improve the accuracy, we set a threshold probability $P_{threshold}$ so that the pixel with low confidence can be excluded. The details of the majority voting strategy can be described by the following formula:

$$L_{mj} = \begin{cases} \text{selected}, & P_{mj} \geq P_{\text{threshold}} \ \& \ L_{mj} = L_{ij} \\ \text{discard}, & \text{else}. \end{cases} \tag{2}$$

In order to improve the robustness of the method, we use Euclidean distance to represent the diversity of samples. Specifically, N samples with the characteristic that the Euclidean distance between any two of them is relatively large are selected for each category from L_{mj}. Those pseudo-labels and the corresponding samples constitute a new pseudo set D_p. We add them to the initial labeled training set D_l and remove them from the unlabeled data set D_u. The process is repeated until the final training set is optimal. Note that, few iterations are enough to obtain a good classification performance.

III. Experimental Results

A. Data Sets

In order to verify the effectiveness of the proposed MCE-SL method, we evaluate it in the following two real HSI data sets: University of Pavia Data set and Botswana Data set.

1) Pavia Data set: The Pavia data set was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor. The image size is $610 \times 340 \times 103$ (where 12 bands with noises and same samples with no information were removed before the analysis), and its wavelength ranging from 0.43 to 0.86 μm. The spatial resolution is up to 1.3-m per pixel. Fig. 2(a) and (b) show the corresponding color image and ground truth, in which nine classes are contained.

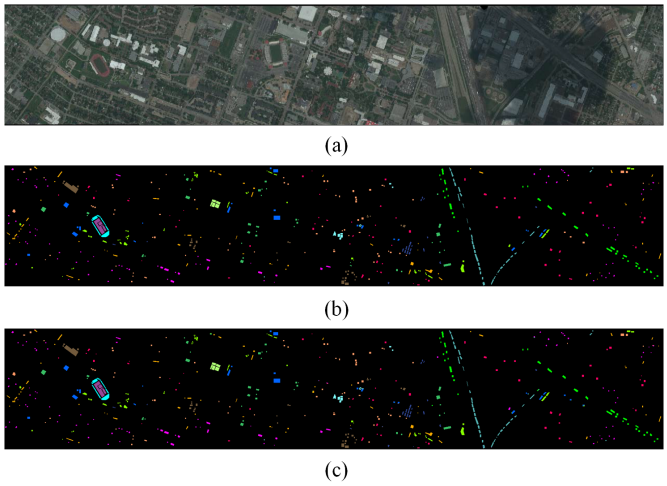

Fig. 3. Houston data set. (a) Color image. (b) Ground truth. (c) Classification map of labeled samples.

2) Houston Data set: Houston data set is a HSI that covers the area of Houston University and its neighboring urban. 15 land cover classes are included in a space of 349×1905 pixels, where the spatial resolution is 2.5 m, and each pixel has 144 bands over a wavelength of $380 \sim 1050$ nm. Fig. 3(a) and (b) show the corresponding color image and ground truth.

B. Experimental Setting

In the proposed MCE-SL method, a VGG like network architecture (3D-CNN) is adopted as a general model. The model includes five convolutional block, two pooling layers and one fully connected layers. A 3-D patch of size $d \times d \times m$ is received as input data of the first CONV layers in order to make full use of the spectral and spatial information, where d is different in submodel, and m indicates the number of spectral bands after the PCA. Each convolutional block contains three CONV layers. The first block has 64 channels, and then the channel is doubled after every pooling process until it becomes 256. The last block is followed by a convolutional layer with a kernel of $d/4$, thereby, the channel of final layer is 512. For the convolutional blocks and pooling layers, their convolutional kernels are 3×3 and 2×2, respectively, and the stride are 1 and 2. The architecture is shown in Table I.

In this experiment, the exponential decay is used in the proposed MCE-SL network. The base learning rate is set as 0.001 and learning rate decay is 0.9. Momentum with 0.99 is used as optimizer. The regularization parameter is 0.0005, and 5000 steps are trained in every iteration. The parameter $P_{threshold}$ is set as 0.6. That value can provide richer diversity in Euclidean distance calculation (z-score normalization is performed in data preprocess) for more pseudo-labels can be included. Besides, the introduced error labels caused by P is also within the tolerance of the model, and do not have a negative effect on the results. The N, representing the maximum number of new selected samples for each category, is set as 30. It should be noted that the actual number of specific selections is automatically calculated by Euclidean distance. In addition, five bands are reserved in the two test images after PCA.

TABLE I
ARCHITECTURE OF THE VGG LIKE NET, WHERE d IS THE SIZE OF PACTH, n IS THE NUMBER OF CLASSES

Layers	Conv block(1-3)	Pooling(4)	Conv block(5-7)	Pooling(8)	Conv block(9-11)	Conv layers(12)	Fc(13)
$Kernel$	3×3	2×2	3×3	2×2	3×3	$d/4$	-
$Step\ length$	1	2	1	2	1	1	-
$Output\ channels$	64	64	128	128	256	512	n

Multiscale patch used as the input of corresponding submodel for the spatial information they contain is different, the diversity will bring on model variance, and it is helpful. The variance mainly expressed as the classification accuracy of the same class is different in different scales, so that the superior performance of one submodel on one class compensates for the other models. As pointed by Haut *et al.* [13], the larger size of spatial patch can characterize more spatial-contextual information around each pixel. In our experiments, patches with sizes of [15, 17, 19, 21, and 23] are used as input data. After the training process, the model with a patch size of 23 is utilized as the final test model since it can obtain more spatial-contextual information, and get a better performance in classification.

C. Experimental Results

To illustrate the performance of the proposed method, the MCE-SL is compared with several traditional methods: SVM [2], deformable HSIs classification networks (DHCNet) [22], and the proposed baseline VGG-like network (3D-CNN). Meanwhile, another two spectral-spatial based semisupervised methods, i.e., the logistic regression via variable splitting and augmented Lagrangian (LORSAL) classifier with AL (LORSAL-AL) [8], RW-based AL and semisupervised learning (SSL) framework (RWASL) [17], are also used as a comparison. In all experiments, the parameters of the comparison method remained unchanged. The overall accuracy (OA), average accuracy (AA), class accuracy (CA) and the Kappa statistic (k) are adopted as the measurement index. All the results listed in Tables II and III are the average values of 5 Monte Carlo runs.

Table II shows the classification accuracy of the comparison experiments on the Pavia data set, in which 50% samples of each class are randomly selected as the testing set. Five samples per class are randomly chosen from another 50% samples, and adopted as the initial training set, while the candidate set consists of the remaining ground truth samples.

The results reported in Table II show that all semisupervised methods including LORSAL-AL and RWASL can get much higher accuracy than the supervised method given the same number of samples, which is attributed to the self-improvement ability of SSL. On the other hand, our method also has obvious advantages over the other two semisupervised methods. On Pavia, the OA increased by 9.97% and 1.2%, respectively. From the above two perspectives, it can be concluded that our method not only effectively solves the problem that requires large amount of data, but also achieves more accurate classification compared with other two methods. Fig. 2(c) shows the classification result obtained from the proposed method.

TABLE II
CLASSIFICATION ACCURACY (%) WITH DIFFERENT METHODS IN PAVIA DATA SET, WHERE FIVE SAMPLES OF EACH CLASS ARE SELECTED RANDOMLY. THE VALUES IN THE PARENTHESES REFER TO THE AVEDEV OF THE ACCURACY

Class	Supervised methods			Semi-supervised methods		
	SVM	DHCNet	3D-CNN	LORSAL-AL	RWASL	MCE-SL
1	91.77(1.82)	93.72(1.57)	82.07(7.76)	88.72(2.51)	99.29(0.88)	99.67(0.35)
2	83.57(2.19)	94.39(1.72)	93.29(4.54)	95.99(1.10)	99.86(0.26)	99.45(0.24)
3	39.04(8.42)	68.67(6.93)	45.48(6.56)	67.58(7.13)	99.63(0.47)	99.98(0.04)
4	52.08(10.1)	63.54(16.0)	55.04(11.1)	87.51(2.05)	81.37(3.71)	99.04(0.65)
5	90.34(9.67)	93.86(3.41)	83.74(9.15)	98.43(0.44)	99.98(0.05)	99.79(0.29)
6	33.32(3.35)	45.72(2.00)	55.43(15.0)	83.66(1.29)	99.29(1.42)	99.57(0.46)
7	43.02(6.21)	47.31(7.99)	33.05(6.64)	69.92(8.40)	99.94(0.19)	99.16(0.66)
8	65.10(5.60)	74.37(4.25)	52.17(17.8)	81.60(5.11)	98.97(1.28)	99.41(0.42)
9	99.85(0.12)	52.04(18.8)	61.57(14.3)	99.51(0.39)	99.33(0.93)	99.26(1.31)
OA	63.51(3.14)	71.88(2.03)	68.93(5.49)	89.52(0.36)	98.29(0.48)	99.49(0.12)
AA	66.46(2.24)	70.40(2.77)	62.43(6.05)	85.88(0.73)	97.52(0.57)	99.49(0.10)
k	55.19(3.19)	65.32(2.15)	60.52(5.29)	86.04(0.45)	97.72(0.65)	99.33(0.16)

TABLE III
CLASSIFICATION ACCURACY (%) WITH DIFFERENT METHODS IN HOUSTON DATA SET, WHERE FIVE SAMPLES OF EACH CLASS ARE SELECTED RANDOMLY. THE VALUES IN THE PARENTHESES REFER TO THE AVEDEV OF THE ACCURACY

Class	Supervised methods			Semi-supervised methods		
	SVM	DHCNet	3D-CNN	LORSAL-AL	RWASL	MCE-SL
1	89.70(6.91)	81.26(4.00)	74.00(6.02)	97.45(1.17)	88.65(5.51)	97.09(2.33)
2	79.79(11.3)	76.41(5.90)	69.50(9.63)	97.85(0.67)	78.10(10.5)	98.41(0.51)
3	70.89(18.1)	86.92(5.34)	92.15(10.1)	99.86(0.10)	100.0(0.00)	100.0(0.00)
4	84.37(12.3)	59.04(4.04)	80.84(5.92)	95.93(1.49)	78.50(5.98)	96.85(2.90)
5	85.64(3.87)	90.28(1.64)	86.18(4.60)	96.49(1.27)	98.94(2.02)	99.52(0.38)
6	83.49(9.45)	63.15(5.09)	85.80(8.21)	93.56(5.09)	90.16(7.17)	93.43(2.63)
7	67.94(11.7)	69.34(7.82)	72.74(7.34)	85.13(2.47)	80.66(4.80)	97.76(1.76)
8	59.41(8.43)	85.59(4.38)	82.28(8.39)	72.87(1.08)	80.61(6.36)	97.47(1.73)
9	59.92(7.18)	67.31(5.50)	67.84(5.71)	79.33(2.96)	79.45(10.9)	96.05(2.01)
10	46.14(9.02)	54.04(2.22)	86.94(4.85)	78.64(7.03)	94.90(6.51)	95.70(2.19)
11	48.91(6.11)	61.08(8.72)	62.12(2.94)	73.92(6.12)	90.09(7.93)	98.67(0.55)
12	41.95(4.98)	66.87(9.27)	75.14(9.39)	80.11(2.89)	92.28(4.52)	98.94(0.80)
13	16.30(4.30)	33.87(4.35)	77.62(10.2)	44.35(7.83)	67.22(7.29)	93.74(2.24)
14	72.32(20.3)	80.27(2.94)	84.97(3.17)	94.75(2.10)	100.0(0.00)	100.0(0.00)
15	99.46(0.27)	87.43(2.77)	84.35(2.56)	98.32(0.32)	100.0(0.00)	100.0(0.00)
OA	65.62(2.30)	68.75(2.04)	76.54(3.00)	86.14(0.85)	87.31(1.51)	97.64(0.23)
AA	67.08(2.58)	70.72(1.69)	78.82(1.90)	85.90(0.93)	87.97(1.35)	97.58(0.29)
k	62.86(2.47)	66.34(2.17)	74.65(3.26)	85.00(0.92)	86.27(1.64)	97.45(0.25)

Fig. 4 describes the changes of OA of the three semisupervised methods with the increase of iterations in Pavia (It should be noted that the RWSAL only presents 10 times but 16 times are iterated actually). It is obvious that the performance of the proposed MCE-SL improves greatly within fewer iterations, and eventually has a higher accuracy. Therefore, the application of multiscale patch do provide richer spatial information, and the majority voting also improves the confidence of pseudo-label.

Table IV shows the number of pseudo-samples required to obtain the classification accuracy in Table II. As can be seen, LORSAL-AL achieved an OA of 89.2% under the

TABLE IV

NUMBER OF PSEUDO-SAMPLES REQUIRED TO OBTAIN THE CLASSIFICATION ACCURACY IN TABLE II

Dataset	Number of Samples		
	LORSAL-AL	RWASL	MCE-SL
Pavia	545	295	2745
Houston	575	325	1875

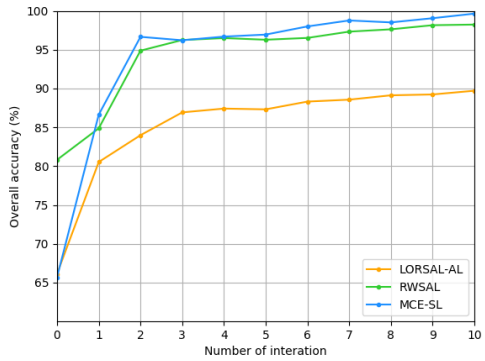

Fig. 4. Effect of number of iterations on the classification accuracy of different methods in Pavia. Where 545, 295, and 2745 new samples are added in end for LORSAL-AL, RWASL, MCE-SL, respectively.

newly added 500 samples. RWASL achieved a good result of 98.29% with less than 300 samples. We increase the number of newly added samples of RWASLs to 2100 (obtained by active-learning and self-learning) within the iterations of 20 times, and the OA is improved to 99.13%. It also can be seen from the table that our method achieves the highest accuracy with the most number of samples. The newly increased samples of the other two methods are manually set, while that in our method is completely automatic. Under the condition that ensuring accuracy, we can obtain more samples in one iteration, which is an advantage of our method.

The size of land cover in Houston is relatively smaller compared with Pavia data set, but our multiscale method still achieves good results. It can be seen from Table III that compared with LORSAL-AL and RWASL, the MCE-SL has a significant improvement in OA, with an increase of 11.5% and 10.33%, respectively. In addition, the average deviation (AVEDEV) of MCE-SL is smaller in all metrics, which reflects the stronger stability of our method. The classification map is shown in Fig. 3(c). Furthermore, our method performs well on two different data sets, indicating that the model has superior generalization.

IV. CONCLUSION

In this letter, the proposed MCE-SL method was introduced. Given a few labeled samples with multiscales spatial information, the trained classifier can automatically select the most useful unlabeled samples to enrich the training set, and promote the behavior of the model progressively. Experiments conducted on two real HSI data sets indicate that the proposed method can achieve higher classification accuracy compared to other well-known classifiers.

REFERENCES

[1] 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.*, vol. 56, no. 6, pp. 3534–3546, Jun. 2018.

[2] 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.

[3] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 8, pp. 492–501, Mar. 2005.

[4] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[5] 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.

[6] N. He *et al.*, "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 755–769, Feb. 2019.

[7] F. Bovolo, L. Bruzzone, and L. Carlin, "A novel technique for subpixel image classification based on support vector machine," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2983–2999, Nov. 2010.

[8] 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.

[9] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Hyperspectral image classification using random occlusion data augmentation," *IEEE Geosci. Remote Sens. Lett.*, to be published.

[10] S. Burr, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin–Madison, Madison, WI, USA, Tech. Rep. 1648, Jan. 2009.

[11] Y. Fu, X. Zhu, and B. Li, "A survey on instance selection for active learning," *Knowl. Inf. Syst.*, vol. 35, no. 2, pp. 249–283, 2013.

[12] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

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

[14] I. Dópido, J. Li, P. R. Marpu, A. Plaza, J. M. B. Dias, and J. A. Benediktsson, "Semisupervised self-learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 4032–4044, Jul. 2013.

[15] D. Tuia and G. Camps-Valls, "Semisupervised remote sensing image classification with cluster kernels," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 224–228, Apr. 2009.

[16] F. de Morsier, M. Borgeaud, V. Gass, J.-P. Thiran, and D. Tuia, "Kernel low-rank and sparse graph for unsupervised and semi-supervised classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3410–3420, Jun. 2016.

[17] B. Sun, X. Kang, S. Li, and J. A. Benediktsson, "Random-walker-based collaborative learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 212–222, Jan. 2017.

[18] X. Kang, S. Li, L. Fang, M. Li, and J. A. Benediktsson, "Extended random walker-based classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 144–153, Jan. 2015.

[19] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.

[20] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 44–57, Jan. 2002.

[21] S. Prasad and L. M. Bruce, "Limitations of principal components analysis for hyperspectral target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 625–629, Oct. 2008.

[22] J. Zhu, L. Fang, and P. Ghamisi, "Deformable convolutional neural networks for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 8, pp. 1254–1258, Aug. 2018.