# Active Deep Feature Extraction for Hyperspectral Image Classification Based on Adversarial Learning

Xue Wang, Kun Tan, *Senior Member, IEEE*, Cen Pan, Jianwei Ding, Zhaoxian Liu, and Bo Han

***Abstract*—The issues of spectral redundancy and limited training samples hinder the widespread application and development of hyperspectral images. In this letter, a novel active deep feature extraction scheme is proposed by incorporating both representative and informative measurement. First, an adversarial autoencoder (AAE) is modified to suit the classification task with deep feature extraction. Dictionary learning and a multivariance and distributional distance (MVDD) measure are then introduced to choose the most valuable candidate training samples, where we use the limited labeled samples to obtain a high classification accuracy. Comparative experiments with the proposed querying strategy were carried out with two hyperspectral datasets. The experimental results obtained with the two datasets demonstrate that the proposed scheme is superior to the others. With this method, the unstable increase in accuracy is eliminated by incorporating both informative and representative measurement.**

***Index Terms*—Active learning, adversarial autoencoder (AAE), hyperspectral image classification, sparse representation.**

## I. Introduction

HYPERSPECTRAL remote sensing data record both rich spectral and spatial information of the scene, which has resulted in great breakthroughs in the field of land-cover monitoring [1]. The classification of remote sensing imagery is a critical prerequisite in a wide range of applications [2]. However, the problems of data redundancy and an insufficient training set are still a barrier to its widespread application and development. Indeed, training a classifier with a strong capability requires a sufficient amount of labeled training data. With the advances in Earth observation technology, more and more land-cover images are now becoming easily accessible. However, obtaining sufficient labeled samples to train a classifier might not be realistic because of the wide scene coverage and the costly field surveying [3]. Classification of hyperspectral (high feature dimension) images can be difficult in the case of limited labeled samples [4]. To tackle these problems, the most valuable samples should be selected, to promote the performance of the classifier [5].

Methods based on the pattern recognition approach were found to have a certain effect in the early development of hyperspectral classification. In this case, dimensionality reduction, such as feature selection, is carried as a preprocessing step [6]. Although this solution can decompose the problem into several controllable subproblems, the optimal solutions of these subproblems cannot converge to a globally optimal solution. Compared with the handcrafted feature extraction methods, deep learning can achieve a superior classification performance [7], [8]. While these deep learning-based methods usually require "big data," which are limited with regard to hyperspectral imagery [9], active learning has been developed to overcome the difficulties in training with a small number of training samples [10].

Active learning, which reduces the cost of labeling, has attracted much interest in the field of remote sensing classification. The principle of an active learning algorithm is to craft a selection strategy to obtain the most valuable instances to label. The querying methods include query-by-committee [11] and uncertainty sampling [12], which are both designed to select the most informative samples. Another querying strategy is to select the samples that are the most representative, which can be achieved by focusing on the cluster centers of unlabeled instances [13].

In this letter, a novel active deep feature extraction scheme is proposed by combining sparse representation and deep feature learning. More specifically, a softmax function is appended to an adversarial autoencoder (AAE) [14] to construct the conditional AAE (CAAE), which is used for deep feature extraction. The encoder is capable of capturing the distribution characteristics of the hyperspectral imagery, and the following softmax layer aims to query the uncertainty of the instances, which can be used for the informative measurement. Dictionary learning is then utilized to select the most representative samples. After this, the new labeled samples are used to retrain the CAAE.

## II. Methodology

The active learning strategy has been widely utilized in hyperspectral image classification. The process of active learning is that the most valuable samples should be selected

Manuscript received 4 March 2022; revised 19 May 2022; accepted 17 June 2022. Date of publication 30 June 2022; date of current version 14 July 2022. This work was supported in part by the Natural Science Foundation of China under Grant 42001350 and Grant 42171335; in part by the Postdoctoral Science Foundation of China under Grant 2021M691016; in part by the Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China, under Grant KLSMNR-202202; and in part by the Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities, under Grant KFKT-2022-07. *(Corresponding author: Kun Tan.)*

Xue Wang and Kun Tan are with the Key Laboratory of Geographic Information Science (Ministry of Education), the School of Geographic Sciences, and the Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai 200241, China (e-mail: tankuncu@gmail.com).

Cen Pan is with the Shanghai Municipal Institute of Surveying and Mapping, Shanghai 200063, China.

Jianwei Ding and Zhaoxian Liu are with the Second Surveying and Mapping Institute of Hebei, Shijiazhuang 050037, China.

Bo Han is with the Institute of Remote Sensing Satellite, China Academy of Space Technology, Beijing 100094, China.

Digital Object Identifier 10.1109/LGRS.2022.3186493

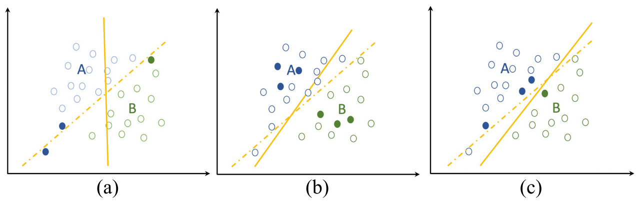

Fig. 1. Three different querying strategies (the points in blue belong to class A, the points in green belong to class B, and the solid points are chosen by the different measurements). (a) Using informative measurement. (b) Using representative measurement. (c) Using combined measurement.

based on a designed rule for the unlabeled dataset, and then, the chosen samples are labeled by experts and used as training samples to retrain the classifier. This process is iterative, and the classifier is trained more effectively than by adding randomly selected new samples each time. Clearly, the selection rule has a significant impact on the final performance, and inappropriate selection can give rise to a waste of effort.

The proposed active deep feature extraction method is based on the integration of representative and informative measurement. Fig. 1 shows the effectiveness of labeling both representative and informative instances for classification. The dashed line in the three figures denotes the optimal classification hyperplane after the classifier training under the assumption that all pixels are labeled in categories A and B. The solid line represents the factual hyperplane based on the actual labeled pixels. Fig. 1(a) shows the instances selected by the informative measurement, but it can be seen that the decision boundaries are incorrect, which is caused by the chosen samples being close to the decision boundaries. Fig. 1(b) shows that the selected instances are capable of identifying the correct decision boundaries, but the improvement is limited because the fine-grained information cannot be added by representative measurement. As shown in Fig. 1(c), the selected samples obtained by the use of both informative and representative measurement are more efficient in finding accurate decision boundaries.

In this work, we adopt a multivariance and distributional distance (MVDD) scheme to choose the most informative samples, and a dictionary learning method is used to find the most representative candidates.

### A. Details of the Proposed Approach

First, the notations adopted throughout this letter are given. If we suppose that a hyperspectral dataset with $b$ spectral bands contains $N$ labeled samples for $L$ classes, and each is represented by $\{x_1, x_2, \ldots, x_N\} \in \mathbb{R}^{1\times b}$, then the corresponding label vector is $Y = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}^{1\times L}$. The last layer of the decoder in the CAAE consists of $b$ neurons, and the generated samples of the CAAE are denoted as $\{x_1^g, x_2^g, \ldots, x_N^g\} \in \mathbb{R}^{1\times b}$.

As shown in Fig. 2, each category has an initial set of labeled samples to train the CAAE, in which all the networks are fully connected. After the training process, the posterior probability matrix is obtained through the softmax layer in the CAAE. Likewise, the generated samples obtained by the trained CAAE are regarded as the items that make up the initial dictionary and are compared with the unlabeled samples. After the calculation of sparse representation coefficient based on orthogonal matching pursuit (OMP) [15], the representative measurement and informative measurement are then applied, and the corresponding candidates are selected in the alternative set, compared with the dictionary items. They are then used to update the dictionary.

### B. Conditional Adversarial Autoencoder

The CAAE is a probabilistic autoencoder that uses an adversarial training process to perform variational inference by matching the aggregated posterior of the hidden code vector with a given prior distribution, which is shown in Fig. 3. The additional softmax is added in AAE to obtain the CAAE, which is capable to tackle the distribution fitting and classification.

Compared with a variational autoencoder, which uses the Kullback–Leibler (KL) divergence to impose a prior distribution on the encoder, the CAAE uses an adversarial training procedure by matching the posterior probability of the hidden encoder with the prior distribution. The output part of the encoder is modified to obtain the posterior probability by appending a softmax layer. The observed loss $L$ consists of three parts—the adversarial loss, the supervised classification loss, and the reconstruction loss—which can be formulated as

$$L = V_{\text{adv}} + \text{Recost} + V_{\text{softmax}}. \tag{1}$$

The solution to the generative adversarial procedure can be expressed as follows [13]:

$$\min_{\text{sampler}} \max_{D} V = \mathbb{E}_{x\sim p(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim p(z)} \times \left[\log(1 - D(\text{sampler}(z)))\right]. \tag{2}$$

The sampler and the discriminator $D$ can be optimized using alternating stochastic gradient descent (SGD): 1) the discriminator is trained to distinguish the true samples from the fake samples generated by the sampler and 2) the generator is then trained to fool the discriminator $D$ with its generated samples. The other two parts of (1) are expressed as

$$(\hat{x} - x)^2 - \sum_{j=1}^{L} y_j \log \frac{e^{a_j}}{\sum_{k=1}^{L} e^{a_k}}. \tag{3}$$

### C. Measurement of Informative Candidates

Dictionary learning based on the deep features is utilized as the representative measurement. The basic assumption of dictionary learning is that the structure of a complex signal can be expressed directly by a set of atoms and the corresponding sparse coefficients, that is, if the sample is used as an atom in a dictionary with high efficiency, it may be the most representative for an active learning problem. The process can be expressed as follows:

$$\max_{D,\alpha_i} \sum_{i=1}^{N} \left\|x_i^h - \text{DIC}\alpha_i\right\|_2^2 + \lambda \sum_{i=1}^{N} \|\alpha_i\|_0 \tag{4}$$

where $N$ denotes the number of items and $\lambda$ is a regularization parameter. $\|\cdot\|_2^2$ and $\|\cdot\|_0$ represent the $L_2$ and $L_0$ norms. With the active learning procedure, DIC is increased, which is solved by the elimination of the residual

$$r_i = x_i^h - \text{DIC}\alpha_i. \tag{5}$$

After this, the sparse coefficients and the items of the dictionary are obtained using OMP and $k$-singular value decomposition in an alternating manner. The iteration is carried out with the updating of the dictionary DIC, using the alternative set chosen by the MVDD scheme.

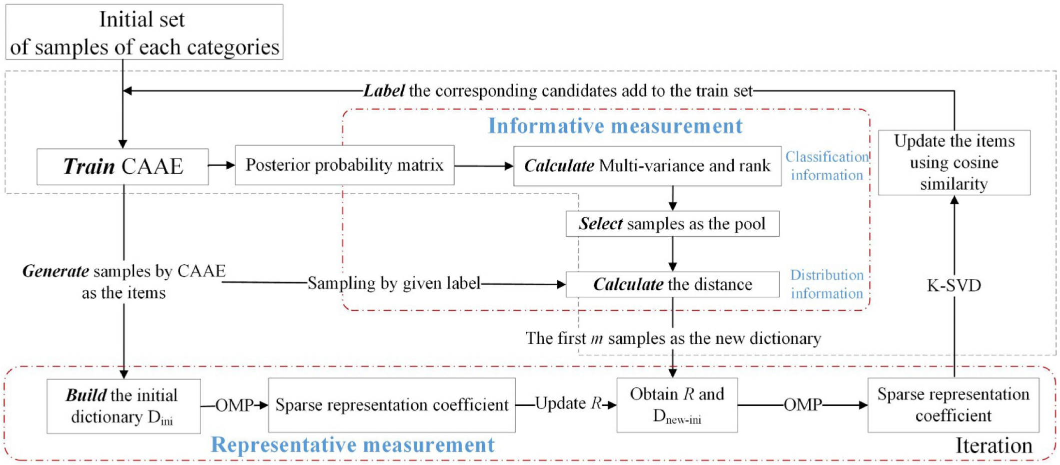
Initial set
of samples of each categories
*Label* the corresponding candidates add to the train set
Informative measurement
*Train* CAAE
Posterior probability matrix
*Calculate* Multi-variance and rank
Classification information
Update the items using cosine similarity
*Select* samples as the pool
*Generate* samples by CAAE as the items
Sampling by given label
*Calculate* the distance
Distribution information
K-SVD
The first *m* samples as the new dictionary
*Build* the initial dictionary D_ini
OMP
Sparse representation coefficient
Update *R*
Obtain *R* and D_new-ini
OMP
Sparse representation coefficient
Representative measurement
Iteration

Fig. 2. Detailed illustration of the proposed MVDD scheme.

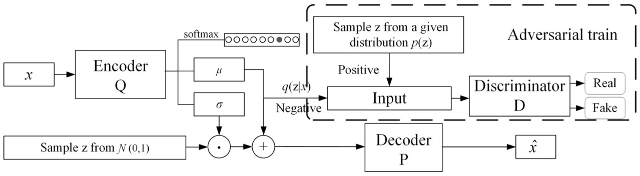

Fig. 3. Structure of the CAAE.

### D. MVDD Scheme

The proposed MVDD scheme consists of two querying strategies: multivariance in the posterior probability and the distance between the learned distribution and the true data distribution. The multivariance calculation uses the top three posterior probabilities. The distributional distance is measured by the use of the Euclidean distance between the generated samples and the real data

$$\begin{aligned}\mathrm{MVDD} = &\max_{i\in L}\left(p(y_l|x) - E(p(y_l|x)\right) + \max_{i\in L\{l^+\}}\left(p(y_l|x)\right.\\ &\left.- E(p(y_l|x)\right) + \max_{i\in L\{l^+,l^{++}\}}\left(p(y_l|x) - E(p(y_l|x)\right)\\ &+ \left\|x^g - x\right\|_2 \end{aligned} \tag{6}$$

where $p(y_l|x)$ means the posterior probability for $y_l$ and $l^+$ is the most probable label class for sample $x$. $l^{++}$ is the second most probable label class for sample $x$. $E(\cdot)$ denotes the expectation of the top three posterior probabilities, and $x^g$ is decoded by the CAAE. A high MVDD value indicates that the corresponding sample has high uncertainty and contains more information.

## III. Experimental Setup and Results

### A. Dataset Description

In order to verify the effectiveness of the proposed method, two hyperspectral datasets were investigated. The first was the Pavia University dataset. This dataset consists of 610 × 340 pixels and is characterized by 103 spectral channels ranging from 430 to 860 nm after noisy band removal. The dataset contains nine categories of interest. Because of the high spatial resolution, the scattered objects, such as the trees and the footpath, bring great difficulty to feature learning. The pseudo-color composite image and the labeled categories are shown in Fig. 4. The second dataset was the Indian Pines dataset. This dataset consists of 145 × 145 pixels, with a spatial resolution of 17 m/pixel. A total of 200 bands ranging from 400 to 2500 nm were used in the experiments after removing the noisy bands. The available training samples cover 16 categories of interest, which are mostly different types of vegetation. The specific ground features of Indian Pines dataset are shown in Fig. 5.

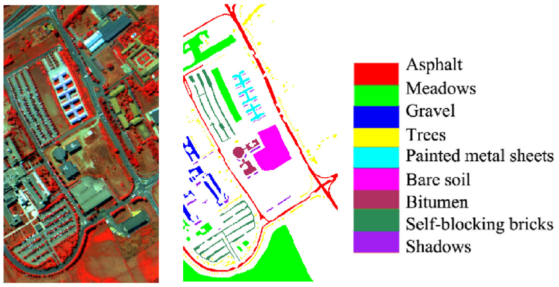

Fig. 4. Pseudo-color composite image and the corresponding ground truth for the Pavia University ROSIS dataset.

### B. Experimental Setting

In order to test the proposed method, the conventional classifier of support vector machine (SVM) and the mainstream deep learning algorithms of the spectral decomposition algorithm (SDA) and deep belief network (DBN) were chosen as comparative methods. SDA and DBN used two layers. Two of the widely used querying strategies were also adopted as the counterpart of the MVDD scheme. Moreover, the two active learning-guided classification models based on deep learning named AL-B-CNN [16] and WI-DL [5] are included as the comparative methods. Because the WI-DL is conducted based on DBN, the network structure of the WI-DL is set consistent with that of DBN.

1) *Maximum Entropy (ME):* It selects instances with the highest classification uncertainty as the most informative samples. The maximum predictive entropy is calculated as

$$\mathrm{ME} = -\sum_{i}^{L} p(y_i|x)\log p(y_i|x). \tag{7}$$

2) *Breaking Ties (BT):* It mainly considers the difference between the largest and the second-largest posterior

TABLE I
NUMBER OF PARAMETERS AND MEMORY OVERHEAD IN ALL METHODS

| Method | | Pavia University dataset | | Indian Pines dataset | |
|---|---|---|---|---|---|
| | | Parameters | Size (MB) | Parameters | Size (MB) |
| SDA | | 60685 | 0.24 | 86420 | 0.34 |
| DBN | | 60194 | 0.23 | 85832 | 0.33 |
| AAE | Q | 30082 | 0.11 | 42498 | 0.16 |
| | P | 31335 | 0.11 | 44744 | 0.17 |
| | D | 17025 | 0.06 | 17025 | 0.06 |
| AL-B-CNN | | 3109 | 0.06 | 5716 | 0.1 |
| WI-DL | | 60194 | 0.23 | 85832 | 0.33 |

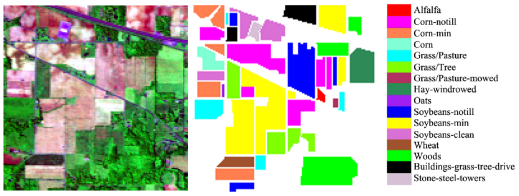

Fig. 5. Pseudo-color composite image and the corresponding ground truth for the Indian Pines AVIRIS dataset.

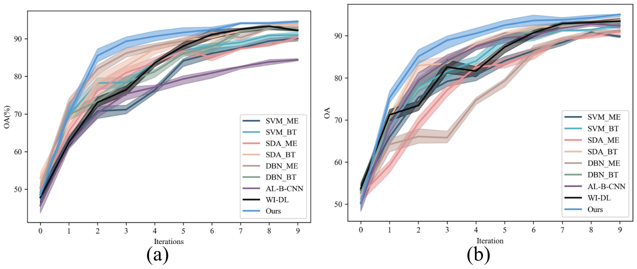

Fig. 6. OAs obtained using the different methods with five initial training samples per class. (a) Indian Pines dataset. (b) Pavia University dataset.

probabilities to measure the similarity between classes, where the smaller the value of BT, the more uncertain the instance is. BT is formulated as

$$\mathrm{BT} = \max_{i \in L} p(y_l|x) - \max_{i \in L\{l^+\}} p(y_l|x) \tag{8}$$

where $l^+ = \mathrm{argmax}_{l \in L}\, p(y_l|x)$ is the most probable label class for sample $x$.

SVM_ME and SVM_BT denote the SVM classifier with ME and BT strategies, respectively. SDA_ME and SDA_BT denote the SDA classifier with ME and BT strategies, respectively. DBN_ME and DBN_BT denote the DBN classifier with ME and BT strategies, respectively. Proposed denotes the proposed method. The hyperparameters of each network were chosen empirically. To exhibit the complexity of the proposed methods, the number of trainable parameters and memory overhead is given in Table I.

### *C. Results*

For the purpose of the comparison, the number of added samples was the same for each active learning algorithm. For a fair comparison, the number of iterations was set to 10, and 50 samples were added in each epoch. The comparative experiments were implemented in six different forms, i.e., the two querying strategies with the three different classification algorithms. The OAs are given in Fig. 6 and Table II, and the classification results are presented in Figs. 7 and 8.

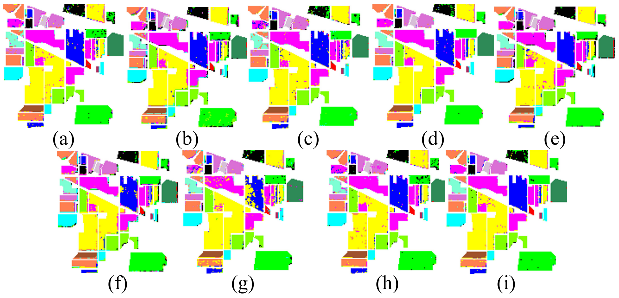

Fig. 7. Classification results for the Indian Pines dataset. (a) SVM_ME. (b) SVM_BT. (c) SDA_ME. (d) SDA_BT. (e) DBN_ME. (f) DBN_BT. (g) AL-B-CNN. (h) WI-DL. (i) Proposed.

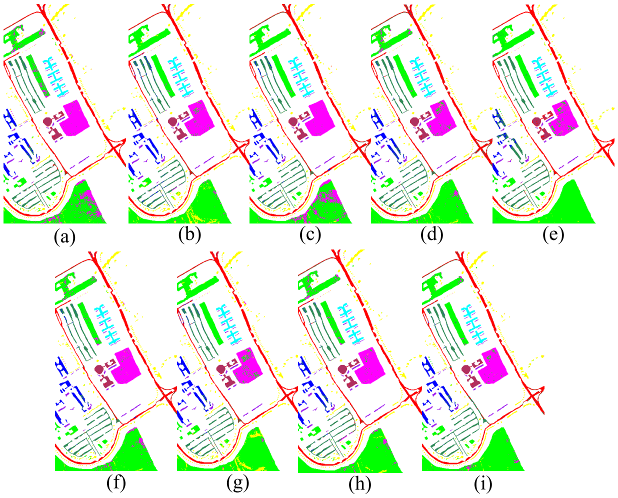

Fig. 8. Classification results for the Pavia University dataset. (a) SVM_ME. (b) SVM_BT. (c) SDA_ME. (d) SDA_BT. (e) DBN_ME. (f) DBN_BT. (g) AL-B-CNN. (h) WI-DL. (i) Proposed.

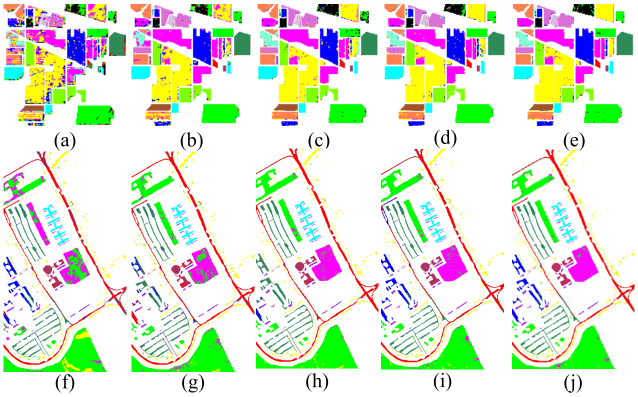

Fig. 9. Classification results of different iterations on the two datasets. (a) IP-Iter1. (b) IP-Iter3. (c) IP-Iter5. (d) IP-Iter7. (e) IP-Iter9. (f) PU-Iter1. (g) PU-Iter3. (h) PU-Iter5. (i) PU-Iter7. (j) PU-Iter9.

The proposed method obtains the best performance of 95.08% and 94.86% in OA on the Pavia University and Indian Pines datasets, respectively, which indicates that the MVDD active learning scheme can significantly improve the classification accuracy when compared with the other querying methods. The OA of the proposed method is better than the other methods by at least 1.32% on the Pavia University dataset and 0.92% on the Indian Pines dataset. Moreover, a decrease in accuracy is seen in both the ME and BT results, which is caused by the unsuitable chosen candidates. The OA obtained by AL-B-CNN has a more stable upward trend with the increment of iteration compared that obtained by WI-DL, while the best OA of WI-DL on the two datasets is higher

TABLE II

OVERALL ACCURACY OF THE DIFFERENT METHODS ON THE PAVIA UNIVERSITY AND INDIAN PINES DATASETS

| Data | Method | Iterations | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Pavia University | SVM_ME | 53.27 | 65.94 | 75.32 | 79.17 | 80.67 | 84.05 | 86.46 | 88.41 | **90.8** | 89.87 |
| | SVM_BT | 53.6 | 68.53 | 77.83 | 81.6 | 83.97 | 89.4 | 90.24 | 91.22 | 91.45 | **92.5** |
| | SDA_ME | 52.13 | 59.23 | 69.24 | 77.09 | 82.51 | 83.32 | 85.68 | 89.04 | 89.93 | **90.97** |
| | SDA_BT | 52.15 | 70.26 | 82.93 | 83.46 | 88.05 | 90.35 | 91.99 | 92.29 | 92.87 | **93.76** |
| | DBN_ME | 51.32 | 64.18 | 66.07 | 65.8 | 74.67 | 79.3 | 86.2 | 88.91 | 90.53 | **91.23** |
| | DBN_BT | 52.67 | 70.2 | 78.34 | 83.7 | 87.38 | 89.77 | 90.89 | 92.5 | 93.1 | **93.09** |
| | AL-B-CNN | 50.2 | 68.72 | 79.38 | 84.05 | 87.43 | 89.49 | 90.1 | 91.83 | **92.77** | 92.39 |
| | WI-DL | 53.7 | 71.21 | 73.44 | 82.56 | 81.61 | 87.33 | 90.74 | 92.89 | 93.14 | **93.49** |
| | Proposed | 50.5 | 75.22 | 85.07 | 88.93 | 90.56 | 92.53 | 93.62 | 93.79 | 94.27 | **95.08** |
| Indian Pines | SVM_ME | 48.73 | 63.04 | 70.75 | 71.18 | 76.32 | 84.15 | 86.14 | 87.6 | 89.42 | **90.11** |
| | SVM_BT | 49.2 | 69.8 | 78.23 | 78.43 | 83.9 | 86.87 | 88.2 | 89.07 | 90.88 | **91.03** |
| | SDA_ME | 50.17 | 65.43 | 76.15 | 80.98 | 83.68 | 86.78 | 85.13 | 87.4 | 88.28 | **90.38** |
| | SDA_BT | 52.89 | 68.22 | 78.4 | 82.7 | 87.33 | 89.7 | 90.32 | 92.1 | 93.51 | **93.94** |
| | DBN_ME | 50.5 | 72.73 | 81.97 | 86.19 | 88.12 | 89.45 | 91.74 | 90.3 | 92.77 | **93.07** |
| | DBN_BT | 48.6 | 69.82 | 73.45 | 78.35 | 81.27 | 86.1 | 87.95 | 91.93 | **92.96** | 92.56 |
| | AL-B-CNN | 45.66 | 61.93 | 70.28 | 75.36 | 77.16 | 79.33 | 80.91 | 82.57 | 83.85 | **84.41** |
| | WI-DL | 47.81 | 62.4 | 73.08 | 76.43 | 83.6 | 88.09 | 91.14 | 92.49 | **93.35** | 92.24 |
| | Proposed | 48.36 | 70.1 | 85.53 | 89.34 | 90.67 | 91.7 | 92.43 | 94.07 | 94.15 | **94.86** |

*The best OA results are marked in bold.

than that of AL-B-CNN. The OA of the proposed method is better than these two additional comparison methods by at least 1.59% on the Pavia University dataset and 1.51% on the Indian Pines dataset. To report the variation of the classification map during the training process, the results of the intermediate process are listed in Fig. 9, which shows that the misclassification phenomenon is modified significantly with the elected samples. Overall, the MVDD scheme eliminates the unstable increase in accuracy by incorporating both informative and representative measurement.

## IV. CONCLUSION

In this letter, a new active learning approach for hyperspectral image classification has been proposed. We also modified the CAAE to make it suitable for the classification task. In this work, we were concerned with the selection of the most informative and most representative unlabeled samples, which is carried out by the MVDD scheme. The proposed MVDD scheme consists of two querying strategies: the multivariance in the posterior probability and the distance between the learned distribution and the true data distribution. Dictionary learning is used as the representative measurement. The proposed method is capable of decreasing the spectral redundancy and highlighting the representative features for the applied task. The experimental results demonstrate that the proposed scheme can result in a significant improvement in classification performance.

## REFERENCES

[1] P. Duan, Z. Xie, X. Kang, and S. Li, "Self-supervised learning-based oil spill detection of hyperspectral images," *Sci. China Technol. Sci.*, vol. 65, pp. 793–801, Mar. 2022.

[2] K. Tan, Y. Zhang, X. Wang, and Y. Chen, "Object-based change detection using multiple classifiers and multi-scale uncertainty analysis," *Remote Sens.*, vol. 11, p. 359, Feb. 2019.

[3] A. J. Brown *et al.*, "Hydrothermal formation of clay-carbonate alteration assemblages in the Nili Fossae region of Mars," *Earth Planet. Sci. Lett.*, vol. 297, nos. 1–2, pp. 174–182, 2010.

[4] X. Wang, K. Tan, Q. Du, Y. Chen, and P. Du, "Caps-TripleGAN: GAN-assisted CapsNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7232–7245, Sep. 2019.

[5] P. Liu, H. Zhang, and K. B. Eom, "Active deep learning for classification of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 712–724, Feb. 2017.

[6] A. Plaza *et al.*, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, Sep. 2009.

[7] Z. Xie, J. Hu, X. Kang, P. Duan, and S. Li, "Multilayer global spectral–spatial attention network for wetland hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518913.

[8] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, "Fusion of dual spatial information for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 7726–7738, Sep. 2020.

[9] V. Singhal and A. Majumdar, "Row-sparse discriminative deep dictionary learning for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 5019–5028, Dec. 2018.

[10] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, Oct. 2014.

[11] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. 5th Annu. Workshop Comput. Learn. Theory (COLT)*, 1992, pp. 287–294.

[12] M.-F. Balcan, A. Broder, and T. Zhang, "Margin based active learning," in *Proc. Int. Conf. Comput. Learn. Theory*. Berlin, Germany: Springer, 2007, pp. 35–50.

[13] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Batch mode active sampling based on marginal probability distribution matching," *ACM Trans. Knowl. Discovery Data*, vol. 7, no. 3, pp. 1–25, Sep. 2013.

[14] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," 2015, *arXiv:1511.05644*.

[15] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[16] 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.