

An End-to-End Joint Unsupervised Learning of Deep Model and Pseudo-Classes for Remote Sensing Scene Representation

Zhiqiang Gong, Ping Zhong, Weidong Hu, Fang Liu, and Bingwei Hui
National Key Laboratory of Science and Technology on ATR
National University of Defense Technology
Changsha 410073, China
E-mail: zhongping@nudt.edu.cn

Abstract—This work develops a novel end-to-end deep unsupervised learning method based on convolutional neural network (CNN) with pseudo-classes for remote sensing scene representation. First, we introduce center points as the centers of the pseudo classes and the training samples can be allocated with pseudo labels based on the center points. Therefore, the CNN model, which is used to extract features from the scenes, can be trained supervised with the pseudo labels. Moreover, a pseudo-center loss is developed to decrease the variance between the samples and the corresponding pseudo center point. The pseudo-center loss is important since it can update both the center points with the training samples and the CNN model with the center points in the training process simultaneously. Finally, joint learning of the pseudo-center loss and the pseudo softmax loss which is formulated with the samples and the pseudo labels is developed for unsupervised remote sensing scene representation to obtain discriminative representations from the scenes. Experiments are conducted over two commonly used remote sensing scene datasets to validate the effectiveness of the proposed method and the experimental results show the superiority of the proposed method when compared with other state-of-the-art methods.

Index Terms—Unsupervised Learning, Pseudo-Class, End-to-End Learning, Convolutional Neural Network (CNN), Remote Sensing Scene Representation

I. INTRODUCTION

Nowadays, high resolution images from the new and the advanced space-borne or aerial-borne sensors contain abundant spatial and spectral information, which could provide helpful information for many military or civilian applications. However, efficient representation and recognition of the remote sensing scenes tend to be a challenging problem since labelling is generally time-consuming and sometimes infeasible [1]. Therefore, unsupervised learning methods tend to be a hot topic to extract discriminative features without the labels of the scenes. The general unsupervised learning methods, such as SIFT [2] and LBP [3], captures the geometrical information, salient points or the textural information from the scenes. However, the complex arrangements in the scenes, the

large inner-class variance and low inter-class variance between different scenes make it difficult to discriminate the scenes from overlapping classes with these low-level features.

In recent years, deep learning methods have shown powerful ability to extract high-level features from the objects. Many deep learning-based unsupervised representations have been developed. It can be divided into four categories: the self-supervised learning approaches which tries to implement a supervised learning with pseudo labels which is created in an unsupervised way, the reconstruction-based methods, the generative adversarial network(GAN)-based methods, and the natural rule motivated loss-function methods [1]. In the first one, the way to construct the pseudo classes plays an important role since they directly affect the training efficiency for the remote sensing scenes. This work will focus on developing an efficient end-to-end unsupervised learning from the self-supervised learning way.

Prior works mainly construct the pseudo classes from three aspects. The first one is to extract the image patches from the scenes and further construct the pseudo classes together with the transformations of the image patches [4]. Some other works take advantage of the selective search method to extract patches with multi-scale information from the scene to form the pseudo classes [5]. Another work makes use of the inner-correlation between the image patches in a scene image to form the pseudo classes [6]. These pseudo classes are generated according to the special requirements of different tasks and the pseudo classes are usually fixed in the training process. However, these prior works mainly take advantage of image patches extracted from the scenes to form the pseudo classes which ignores some important information in the scene image. This would limit the performance of the learned model to extract discriminative features from the scenes.

To overcome this problem, this work denotes the center points to represent the pseudo classes. Based on the center points, we allocate the training samples to the nearest pseudo class. Motivated by the prior work [7], a novel pseudo center loss is formulated with the training samples and the corresponding center point the samples belong to. The center

This work was supported in part by the NSF of China under Grant 61671456 and Grant 61271439, FANEDD under Grant 201243, and Program for New Century Excellent Talents in University under Grant NECT-13-0164.

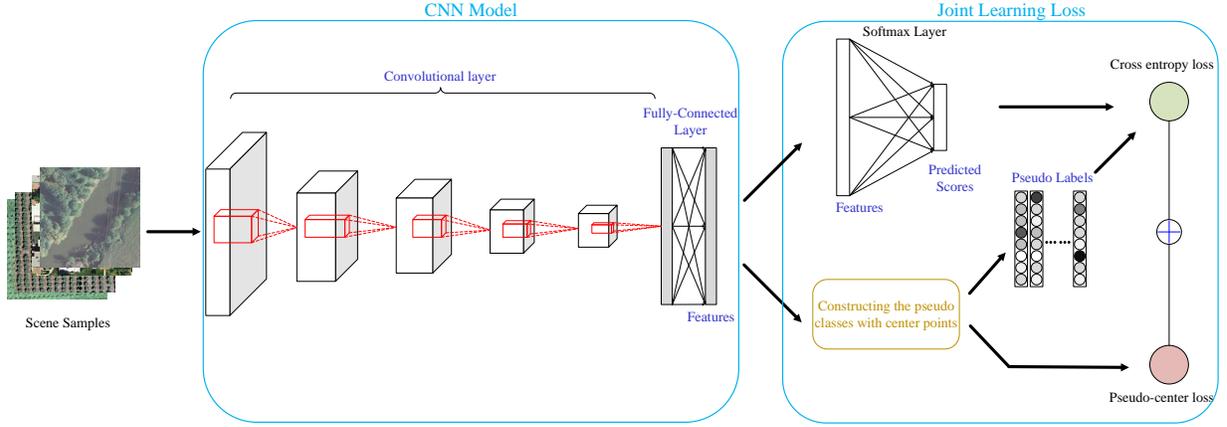


Fig. 1. Flowchart of the proposed method for unsupervised learning of remote sensing scenes. The CNN model is used to extract deep features from the remote sensing scenes. The joint learning loss tries to update both the center points and the CNN model simultaneously. Therefore, the pseudo classes are updated with the update of the center points and the center points tend to describe the centers of different classes. Then, the learned model can be better fit for the remote sensing scenes and can discriminate scenes from different classes.

points would be updated with the pseudo center loss in the training process and the pseudo label of each training sample would also be changed with the updated center points in the unsupervised training process.

To take advantage of both the deep representation and the pseudo center loss, this work develops a novel end-to-end joint unsupervised learning of deep model and pseudo classes for the remote sensing scenes, which jointly learns the pseudo center loss and the pseudo softmax loss. The pseudo softmax loss which is formulated with the pseudo labels is used to update the deep model. Through updating the deep CNN model and the center points which represent the pseudo classes simultaneously, the pseudo classes would be close to the real classes and the learned features from the deep model would be more discriminative.

The rest of the paper is arranged as follows. In section II, we briefly introduce the general convolutional neural network, develop the end-to-end deep unsupervised learning method with pseudo-classes for remote sensing scene representation, and introduces the implementation of the proposed method in detail. Details of our experiments and results are presented in Section III. Section IV concludes the paper with some discussions.

II. PROPOSED METHOD

In this section, we first briefly introduce the general convolutional neural network (CNN), and then develops the pseudo center loss with the pseudo-classes for unsupervised learning of remote sensing scenes, and then the joint learning method for unsupervised learning is developed, and finally we present the implementation of the proposed method for remote sensing scene representation.

Let us denote $\mathbf{x}_i (i = 1, 2, \dots, N)$ as the samples from the remote sensing scenes and N is the number of the unlabelled scenes.

A. General Convolutional Neural Network (CNN)

Deep learning-based method, such as Convolutional Neural Network (CNN), Deep Belief Network (DBN), have shown their impressive performance for remote sensing scene representation [8]. Among these methods, CNNs which can extract both the local and global features from the scenes have been widely used in the literature of remote sensing [8], [9]. As Fig. 1 shows, this work chooses the CNN model to extract features from the scenes.

The general CNNs consist of layers of many types, such as the convolutional layer, fully connected layer, pooling layer, ReLU layer, loss layer. It can be looked as the parallel of these layers where the output of the former layer is performed as the input of the current layer. Denote s^k as the features learned from the k^{th} layer, and then the features s^{k+1} that obtained from the $k + 1^{th}$ layer can be calculated by

$$s^{k+1} = f(W_k s^k + \mathbf{b}_k) \quad (1)$$

where W_k and \mathbf{b}_k represent the parameters and the bias in the k^{th} layer. $f(\cdot)$ denotes the nonlinear activation function.

To accurately train the deep model, the training batch, which denotes a set of samples that train the deep model simultaneously, is usually used in the training process. In addition, the softmax loss, which consists of softmax layer and cross entropy loss, is generally used for the training of the CNN model.

B. Pseudo Center Loss with Pseudo-Classes

Denote $\mathbf{c}_i (i = 1, 2, \dots, \Lambda)$ as the center point where each center point represents a pseudo class and Λ represents the number of the pseudo classes. To provide the pseudo labels to the unlabelled samples, the key process is to formulate the variance between the samples and different pseudo classes.

Given a training batch B . For each sample $\mathbf{x}_i \in B$, denote $\varphi(\mathbf{x}_i)$ as the features extracted from the CNN model. Since the center points $\mathbf{c}_i (i = 1, 2, \dots, \Lambda)$ are constructed to represent

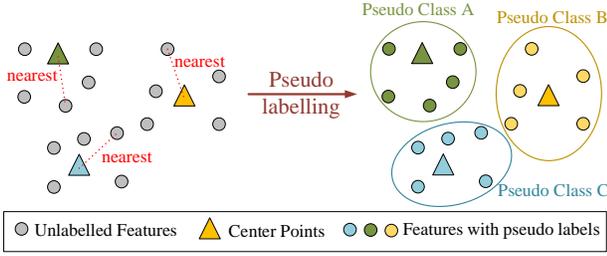


Fig. 2. The process of formulating the pseudo classes with the training samples in the training batch based on the center points. The sample is allocated with the pseudo label from the nearest center point. It should be noted that the pseudo classes would be dynamic changed with the update of the center points and the CNN model.

different classes, the pseudo class each sample in the training batch B belongs to can be calculated by

$$z_i = \arg \min_l \|\mathbf{c}_l - \varphi(\mathbf{x}_i)\|, (l \in \{1, 2, \dots, \Lambda\}), \quad (2)$$

The process for allocating pseudo labels to the training samples in the batch is shown in Fig. 2. Since the pseudo classes and the CNN model are dynamic changed, the pseudo label of each sample is changed in the training process.

Since the construction of the pseudo classes is related to the center points, the update of the center points can significantly affect the effectiveness of the training process as well as the performance of the representation for the scenes. Motivated by [7], this work formulates the pseudo center loss with the samples in the batch to update the center points. The pseudo center loss tries to encourage the center points to approach the training samples in the pseudo class and it can be formulated as

$$L_c = \sum_{i=1}^{|B|} \|\mathbf{c}_{z_i} - \varphi(\mathbf{x}_i)\|^2 \quad (3)$$

where z_i is the pseudo label of \mathbf{x}_i calculated from Eq. 2. In the training process, the L_c is used to update both the parameters in the CNN model and the center points.

C. Joint Learning Loss for Unsupervised Learning of Remote Sensing Scene Representation

As general self-supervised learning approaches, this work uses the softmax loss to supervised learn the CNN model with the pseudo classes. The pseudo softmax loss formulated by the samples in different pseudo classes can be calculated by

$$L_s = - \sum_{i=1}^{|B|} \log \frac{e^{W_{0,z_i}^T \varphi(\mathbf{x}_i) + b_{0,z_i}}}{\sum_{j=1}^{\Lambda} e^{W_{0,j}^T \varphi(\mathbf{x}_i) + b_{0,j}}} \quad (4)$$

where $W_0 = [W_{0,1}, W_{0,2}, \dots, W_{0,\Lambda}]$, $\mathbf{b}_0 = [b_{0,1}, b_{0,2}, \dots, b_{0,\Lambda}]$ represent the parameters and the bias term in Softmax layer, respectively. The L_s is used to calculate the penalization between the predicted scores over the pseudo classes with the pseudo labels by Eq. 2.

Considering the merits of the CNN model and the center points which are used to formulate the pseudo classes, this

work develops a novel joint learning loss of the pseudo classes and the deep model. It can be formulated as

$$\begin{aligned} L &= L_s + \lambda L_c \\ &= - \sum_{i=1}^{|B|} \log \frac{e^{W_{0,z_i}^T \varphi(\mathbf{x}_i) + b_{0,z_i}}}{\sum_{j=1}^{\Lambda} e^{W_{0,j}^T \varphi(\mathbf{x}_i) + b_{0,j}}} + \lambda \sum_{i=1}^{|B|} \|\mathbf{c}_{z_i} - \varphi(\mathbf{x}_i)\|^2 \end{aligned} \quad (5)$$

where λ is a positive value which denotes the tradeoff between the pseudo softmax loss and the pseudo center loss. The developed pseudo classes-based loss can jointly learn the deep model and the pseudo classes simultaneously. Therefore, the learned model can be more fit for the remote sensing scenes and could discriminate scenes with great similarity from different classes.

D. Implementation of the Proposed Method

The proposed unsupervised learning process can be trained end-to-end by the stochastic gradient descent (SGD). According to the characteristics of the back propagation of the deep model [11], the main problem is to calculate the partial of the joint learning loss w.r.t. \mathbf{x}_i . More importantly, in this work, the update of the pseudo classes should also be implemented by calculating the partial of the joint learning loss w.r.t. the center points.

The partial of the pseudo softmax loss L_s w.r.t. \mathbf{x}_i can be calculated as Caffe which is the deep learning framework used in the experiments [10]. The partial of the pseudo center loss can be calculated as [7] shows. Therefore, the partial of the proposed joint learning loss w.r.t. \mathbf{x}_i can be calculated by

$$\frac{\partial L}{\partial \varphi(\mathbf{x}_i)} = \frac{\partial L_s}{\partial \varphi(\mathbf{x}_i)} + 2\lambda(\varphi(\mathbf{x}_i) - \mathbf{c}_{z_i}), \quad (6)$$

where z_i is the pseudo label of \mathbf{x}_i . This is used for the update of the parameters in the CNN model.

In addition, the partial of the proposed joint learning loss w.r.t. \mathbf{c}_j can be calculated as

$$\frac{\partial L}{\partial \mathbf{c}_j} = 2\lambda \sum_{\mathbf{x}_i \in B} I(z_i = j)(\mathbf{c}_j - \varphi(\mathbf{x}_i)). \quad (7)$$

where $I(\cdot)$ represents the indicative function. $\frac{\partial L}{\partial \mathbf{c}_j}$, which is used to update the center points in the training process, can adjust the pseudo classes to the real one and make the learned features from the scenes be discriminative. The overall unsupervised learning framework is given in Algorithm 1.

III. EXPERIMENTAL RESULTS

A. Experimental Setup

To further validate the effectiveness of the proposed method, we conduct experiments over the Ucmcered Land Use dataset [13] and the Brazilian Coffee Scene dataset [8]. The Ucmcered Land Use dataset consists of 2100 high resolution aerial scenes (1 foot per pixel) with 256×256 pixels which can be divided into 21 classes. The Brazilian Coffee Scene dataset contains 2876 multi-spectral scenes with 64×64 pixels which can be divided into 2 classes. Fig. 3 and 4 show the samples from the

Algorithm 1 Implementation of the unsupervised learning method

Require: $\mathbf{x}_i (i = 1, 2, \dots, N)$, $\theta_k = \{W_k, \mathbf{b}_k\}$ as the parameter of the k^{th} convolutional layer, W_0 as the parameters and \mathbf{b}_0 is the bias term in Softmax layer, hyperparameter λ , learning rate lr , the number of pseudo classes Λ .

Ensure: θ_k

- 1: Initialize θ_k in k^{th} convolution layer where W_k is initialized from Gaussian distribution with standard deviation of 0.01 and b_k is set to 0. Initialize the center point $\mathbf{c}_i (i = 1, 2, \dots, \Lambda)$ where \mathbf{c}_i is filled with 0.
- 2: **while** not converge **do**
- 3: $t \leftarrow t + 1$.
- 4: Construct the training batch B^t .
- 5: Obtain the features $\varphi(\mathbf{x}_i^t)$ of $\mathbf{x}_i^t \in B^t$ from CNN model with θ_k^t .
- 6: Obtain the pseudo label z_i^t of $\mathbf{x}_i^t \in B^t$ as Eq. 2 shows.
- 7: Compute the pseudo center loss with the pseudo labels of samples by $L_c^t = \sum_{i=1}^{|B^t|} \|\mathbf{c}_{z_i^t}^t - \varphi(\mathbf{x}_i^t)\|^2$.
- 8: Compute the joint learning loss by $L^t = L_s^t + \lambda L_c^t$ where L_s^t is calculated as Eq. 4.
- 9: Compute the deviation L^t w.r.t. $\varphi(\mathbf{x}_i^t)$ in B^t by $\frac{\partial L^t}{\partial \varphi(\mathbf{x}_i^t)} = \frac{\partial L_s^t}{\partial \varphi(\mathbf{x}_i^t)} + 2\lambda(\varphi(\mathbf{x}_i^t) - \mathbf{c}_{z_i^t}^t)$.
- 10: Compute the deviation L^t w.r.t. \mathbf{c}_j^t by $\frac{\partial L^t}{\partial \mathbf{c}_j^t} = 2\lambda \sum_{\mathbf{x}_i^t \in B^t} I(z_i^t = j)(\mathbf{c}_j^t - \varphi(\mathbf{x}_i^t))$.
- 11: Update the parameters W by $W^{t+1} = W^t - lr \times \frac{\partial L^t}{\partial W^t} = W^t - lr \times \frac{\partial L_s^t}{\partial W^t}$.
- 12: Update the parameters θ_k of k^{th} layer by $\theta_k^{t+1} = \theta_k^t - lr \times \frac{\partial L^t}{\partial \theta_k^t} = \theta_k^t - lr \times \sum_{i=1}^{|B^t|} \frac{\partial L^t}{\partial \varphi(\mathbf{x}_i^t)} \times \frac{\partial \varphi(\mathbf{x}_i^t)}{\partial \theta_k^t}$.
- 13: Update the center points \mathbf{c}_j by $\mathbf{c}_j^{t+1} = \mathbf{c}_j^t - lr \times \frac{\partial L_c^t}{\partial \mathbf{c}_j^t}$.
- 14: **end while**
- 15: **return** θ_k

Ucmerced Land Use dataset and the Brazilian Coffee Scene dataset, respectively.

The deep model is implemented on Caffe which is a commonly used deep learning framework (see [10] for details). CaffeNet is chosen as the deep CNN model to extract unsupervised features from the remote sensing scenes. It should be noted that in the experiments, the dimension of the last fully-connected layer is set to 512 to decrease the parameters in the model and accelerate the training process. In addition, the learning rate, the training epoch are set to 0.00001, 10000, respectively.

With the proposed method, feature vectors of the scene images are obtained. To evaluate the performance of the obtained features, we choose SVM classifier to predict scene

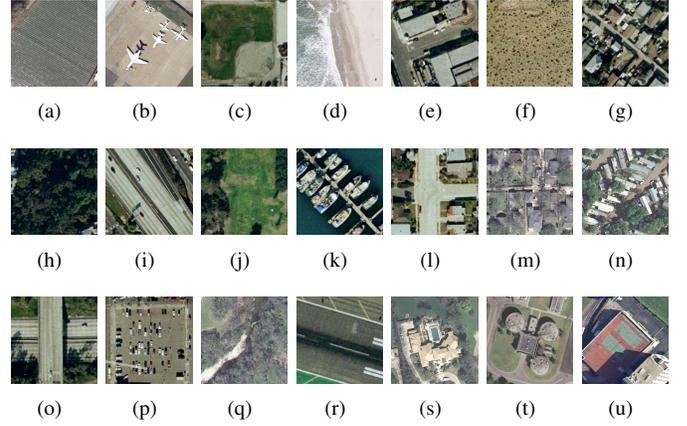


Fig. 3. Scene samples from Ucmerced Land Use dataset. (a) agricultural; (b) airplane; (c) baseball diamond; (d) beach; (e) buildings; (f) chaparral; (g) dense residential; (h) forest; (i) freeway; (j) golf course; (k) harbor; (l) intersection; (m) medium density residential; (n) mobile home park; (o) overpass; (p) parking lot; (q) river; (r) runway; (s) sparse residential; (t) storage tanks; (u) tennis court.

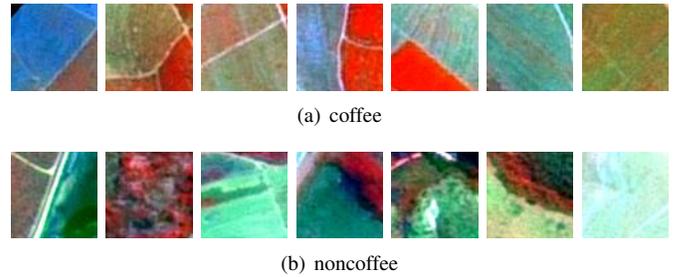


Fig. 4. Samples of different classes from Brazilian Coffee Scene dataset.

labels with the obtained features, and LSSVM is adopted [12]. In the experiments, both the datasets have been equally divided into five folds. To accurately validate the performance of the proposed method, all the results are obtained from the average and the standard deviation of the five-fold cross-validation.

B. Results Over the Ucmerced Land Use dataset

Through experiments over the Ucmerced Land Use dataset, the classification accuracy can achieve 94.33% with the proposed method. The confusion matrix can be seen in Fig. 5. From the confusion matrix, we can find that only some classes with great similarity could not be separated with the proposed method, such as the denseresidential and the mediumresidential, the mediumresidential and the sparseresidential. The classification errors of denseresidential/mediumresidential and mediumresidential/denseresidential can be 10% and the error of sparseresidential/mediumresidential is 5%. Most of the classes can be discriminated. The results show the effectiveness of the proposed method for unsupervised learning of remote sensing scenes. However, the classification performance of the proposed method can be affected by the hyper-parameter λ and the number of pseudo classes.

1) *Classification Performance with Different Hyperparameter λ* : As subsection II-C shows, the λ denotes the

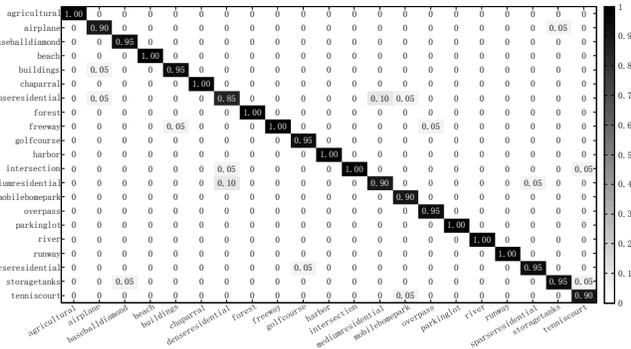


Fig. 5. Confusion matrix of the proposed method over the Ucmcerced Land Use dataset.

tradeoff between the pseudo softmax loss and the pseudo center loss. The pseudo center loss have significant effects on the update of the center point of each pseudo class and therefore the classification performance can be significantly affected by the hyper-parameter λ .

Fig. 6 presents the classification performance of the proposed method with different λ over Ucmcerced Land Use. The results are obtained when the pseudo classes is set to 10. We can find from the tendency of accuracies with different λ in Fig. 6 that with the increase of the value of λ , the center points of the pseudo classes can be fully learned and be more accurate to describe the unlabelled data. Therefore, the classification performance is improved. However, when the lambda is extensively large, the training process focuses too much attention on the update of the center points which may cause the decrease of the classification performance. In addition, from Fig. 6, we can find that the performance can obtain $94.33\% \pm 1.06\%$ which ranks the best when λ is set to 10^{-5} .

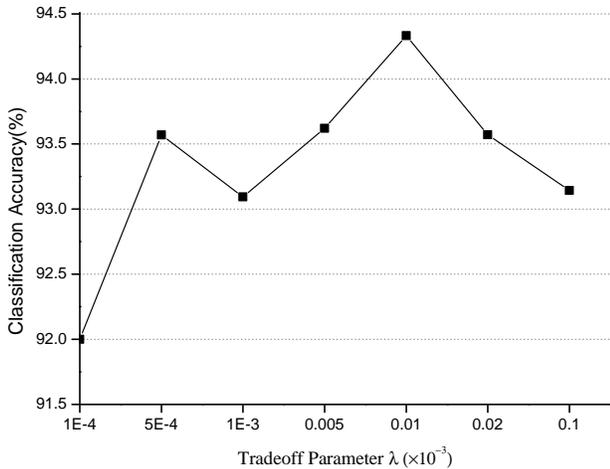


Fig. 6. Classification accuracy obtained by the proposed method with different tradeoff parameter λ over Ucmcerced Land Use dataset.

2) *Classification Performance with Different Number of Pseudo-Classes:* In the experiments, we choose 2, 5, 10, 15, 21 as the number of pseudo-classes for the proposed method

over the Ucmcerced Land Use dataset, respectively. The number of pseudo classes has obviously effects on the classification performance of the proposed method.

Fig. 7 shows the classification accuracies of the proposed method with different number of pseudo classes over Ucmcerced Land Use dataset. In the experiments, the hyper-parameter λ is set to 10^{-4} . We can find that with the increase of the pseudo classes, the classification performance is improved and too many pseudo classes would decrease the performance. Too small pseudo classes would make samples from different classes be assigned to the same class, and therefore the learned model could not separate different samples. In contrast, too many pseudo classes would make samples from the same class be assigned to different pseudo classes, which would also decrease the classification performance. From Fig. 7, it can be noted that the classification performance can achieve 94.33% when the number of pseudo-classes is set to 10.

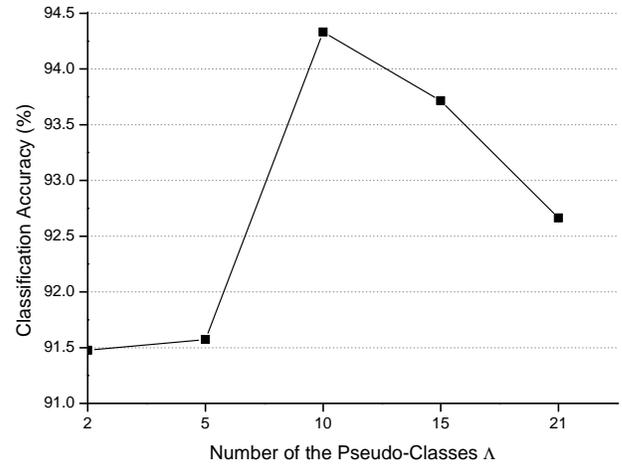


Fig. 7. Effects of the number of pseudo-classes on the performance of the proposed method over Ucmcerced Land Use dataset.

3) *Comparisons with the Most Recent Methods:* To comprehensively validate the effectiveness of the proposed method for unsupervised learning of remote sensing scene representation, we compare the proposed method with other state-of-the-art methods. Table I lists the classification accuracies of several state-of-the-art unsupervised learning methods over Ucmcerced Land Use dataset. From the table, we can find that the proposed method which can obtain 94.33% outperforms other hand-crafted features, such as Dense SIFT (81.67%) [14], SPCK++ (76.05%) [15], UFL-SC (90.26%) [16], and COPD (91.33%) [17]. In addition, when compared with other deep models, such as MARTA GANs without data augmentation (85.37%) [1], [18], CNN-1 (84.53%) [19] and UCFFN (88.57%) [1], the proposed method can also obtain better performance. Therefore, the proposed method can obtain comparable or even better performance over the Ucmcerced Land Use dataset when compared with other state-of-the-art methods, including the hand-crafted feature-based methods, and deep methods.

TABLE I
COMPARISONS WITH OTHER RECENT METHODS FOR UNSUPERVISED
LEARNING OVER UC MERCED LAND USE DATASET.

Methods	Accuracy(%)
Dense SIFT [14]	81.67 ± 1.23
SPCK++ [15]	76.05
UFL-SC [16]	90.26 ± 1.51
COPD [17]	91.33 ± 1.11
MARTA GANs (without data augmentation) [1], [18]	85.37
CNN-1 [19]	84.53
UCFFN [1]	88.57
Proposed Method	94.33 ± 1.06

C. Results Over the Brazilian Coffee Scene dataset

The proposed method can obtain 87.74% + 1.59% over Brazilian Coffee Scene dataset. The corresponding confusion matrix is shown in Fig. 8. From the confusion matrix, we can find that the classification errors of coffee/noncoffee, and noncoffee/coffee are 9% and 11%, respectively. Obviously, the classification performance of the proposed method can be significantly affected by λ and the number of pseudo classes.

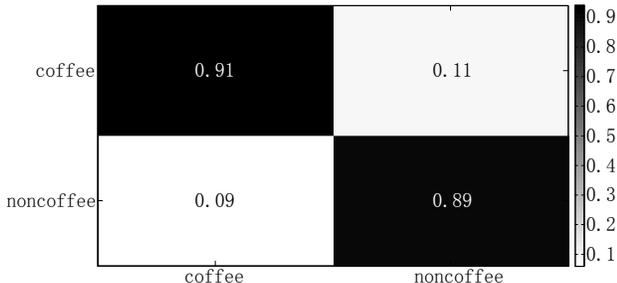


Fig. 8. Confusion matrix of the proposed method over the Brazilian Coffee Scene dataset.

1) *Classification Performance with Different Hyper-parameter λ* : Over Brazilian Scene dataset, we also investigate the results with the sets of λ as $\{10^{-7}, 5 \times 10^{-7}, 10^{-6}, 5 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}, 1 \times 10^{-4}\}$. The classification results with different λ are shown in Fig. 9. From the tendency, we can find that it is important to choose a proper λ for the proposed method and the proposed method achieves 87.74% which ranks the best when λ is set to 10^{-5} . It should also be noted that when λ is extensively large, the training process may not be converged.

2) *Classification Performance with Different Number of Pseudo-Classes*: Just as the Ucmcerced Land Use dataset, the number of the pseudo classes can affect the classification performance. Since the Brazilian Coffee Scene dataset contains two classes, this work conducts experiments over the Brazilian Coffee Scene dataset where the number of pseudo-classes is chosen from $\{2,3,4,5,6,7,8\}$.

Fig. 10 shows the classification performance of the proposed method with different number of pseudo-classes with the hyper-parameter $\lambda = 1 \times 10^{-5}$. We can find from the figure that the proposed method achieves 87.74% which ranks the best when the number of pseudo classes is set to 5.

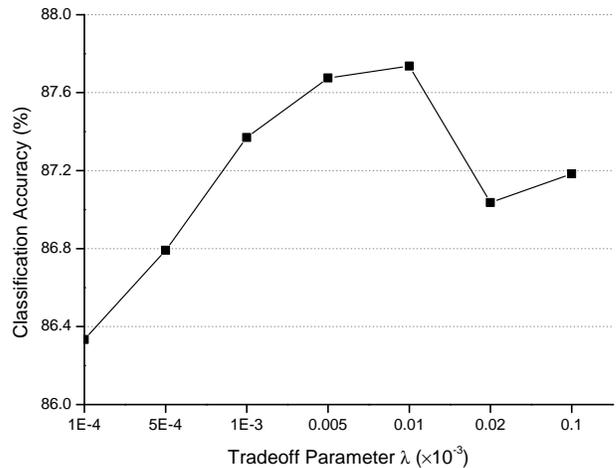


Fig. 9. Classification accuracy obtained by the proposed method with different tradeoff parameter λ over Brazilian Coffee Scene dataset.

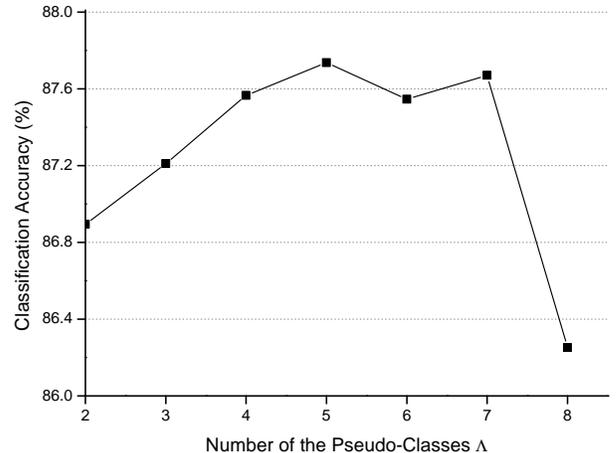


Fig. 10. Effects of the number of pseudo-classes on the performance of the proposed method over Brazilian Coffee Scene dataset.

3) *Comparisons with the Most Recent Methods*: Table II lists the classification accuracies of several state-of-the-art methods over the Brazilian Coffee Scene dataset. The proposed method which can obtain 87.74% outperforms other shallow methods, such as SIFT (82.83%) [9], BIC (87.03%) [8], BOVW (80.50%) [8], and OverFeat_L+OverFeat_S (83.04%) [8]. When compared with other deep models, the proposed method can obtain comparable results. The proposed method can obtain 87.74% which is better than 86% obtained by CNN-1 [19] and 87.69% by MARTA GANs without data augmentation [1], [18]. It can obtain comparable results when compared with UCFFN (87.83%). The comparisons show the superiority of the proposed method on unsupervised learning of the remote sensing scenes.

IV. CONCLUSIONS

This work proposes a novel end-to-end unsupervised learning method for the representation of remote sensing scenes. First, the proposed method chooses the CNN model to extract

TABLE II
COMPARISONS WITH OTHER RECENT METHODS FOR UNSUPERVISED
LEARNING OVER BRAZILIAN COFFEE DATASET.

Methods	Accuracy(%)
SIFT [9]	82.83
BIC [8]	87.03 \pm 1.07
BOVW [8]	80.50
OverFeat _L +OverFeat _S [8]	83.04 \pm 2.00
CNN-1 [19]	86.00
MARTA GANs (without data augmentation) [1], [18]	87.69
UCFFN [1]	87.83
Proposed Method	87.74 \pm 1.59

features from the scenes. Then, center points are introduced in the training process to formulate the pseudo classes. By allocating pseudo labels to different samples based on the center points and formulating the pseudo center loss to decrease the variance between the samples and the corresponding center point, different samples can be clustered with the center points. In addition, through joint learning of the pseudo center loss and the pseudo softmax loss, the center points and the parameters in CNN model are both updated. Experimental results show that the proposed end-to-end unsupervised learning process can extract discriminative features from the scenes. In addition, the proposed method can obtain comparable or even better results when compared with other state-of-the-art methods.

In future work, we intend to apply the proposed unsupervised learning methods on other computer vision tasks, such as visual representation. In addition, we also would like to evaluate the performance of the proposed method with other CNN model, such as GoogLeNet, ResNet. Moreover, other technologies, which can improve the performance of the unsupervised learning methods, would be another interesting topic.

REFERENCES

- [1] Y. Yu, Z. Gong, C. Wang, and P. Zhong, "An Unsupervised Convolutional Feature Fusion Network for Deep Representation of Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 1, pp. 23–27, 2018.
- [2] S. Chen and Y. Tian, "Pyramid of spatial relations for scene-level land use classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 1947–1957, 2015.
- [3] J. Ren, X. Jiang, and J. Yuan, "Learning LBP structure by maximizing the conditional mutual information," *Pattern Recognition*, vol. 48, no. 10, pp. 3180–3190, 2015.
- [4] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1734–1747, 2016.
- [5] A. Ghaderi and V. Athitsos, "Selective unsupervised feature learning with convolutional neural network (S-CNN)," *arXiv preprint arXiv:1606.02210*, 2016.
- [6] M. Noroozi and P. Favaro, "Unsupervised Learning of visual representations by solving jigsaw puzzles," *In European Conference on Computer Vision*, pp. 69–84, 2016.
- [7] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," *In European Conference on Computer Vision*, pp. 499–515, 2016.

- [8] O. A. B. Penati, K. Nogueira, and J. A. d. Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 44–51, 2015.
- [9] Z. Gong, P. Zhong, Y. Yu, and W. Hu, "Diversity-Promoting deep structural metric learning for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 371–390, 2018.
- [10] Y. Jia et al, "Caffe: Convolutional architecture for fast feature embedding," *In proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678, 2014.
- [11] S. S. Haykin, "Neural Networks and Learning Machines," UpperSaddle River: Pearson, 2009.
- [12] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.
- [13] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," *In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279, 2010.
- [14] A. M. Cheryadat, "Unsupervised feature learning for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 439–451, 2014.
- [15] Y. Yi and S. Newsam, "Spatial pyramid co-occurrence for image classification," *IEEE International Conference on Computer Vision*, pp. 1465–1472, 2011.
- [16] F. Hu, G. S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 5, pp. 2015–2030, 2015.
- [17] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 119–132, 2014.
- [18] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: Unsupervised representation learning for remote sensing image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2092–2096, 2017.
- [19] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1349–1362, 2016.