Cross-Dataset Hyperspectral Image Classification Based on Adversarial Domain Adaptation

Xiaorui Ma, *Member, IEEE*, Xuerong Mou, Jie Wang, *Senior Member, IEEE*, Xiaokai Liu, Jie Geng, *Member, IEEE*, and Hongyu Wang, *Member, IEEE*

***Abstract*— The cross-data set knowledge is vital for hyperspectral image classification, which can reduce the dependence on the sample quantity by transferring knowledge from other data sets and improve the training efficiency by sharing knowledge between different data sets. However, due to the capturing environment change and imaging equipment difference, domain shift troubles the exploitation of the cross-data set knowledge. To address the aforementioned issue, this article proposes an unsupervised cross-data set hyperspectral image classification method based on adversarial domain adaptation. The proposed method, which employs multiple classifiers to build a discriminator and uses variational autoencoders to constitute a generator, works in an adversarial manner to drive the target samples under the support of the source domain. In particular, the classification error and the classification disagreement are considered in the objective function, which helps to align different domains while keeping the boundaries of different classes. Experimental results of the multidomain data set demonstrate that the proposed method can transfer and share cross-data set knowledge and achieve state-of-the-art performance without using the labeled information of the target data set.**

***Index Terms*— Classification, cross-data set, domain adaptation, hyperspectral image.**

I. Introduction

HYPERSPECTRAL images, which are usually mounted on remote sensing platforms, combine the advantages of the spectrograph and optical cameras and provide hyperspectral images with both spectral information and spatial information of the objective scene [1], [2]. The high-resolution spectral information and the large-scale spatial information afford to achieve fine-grained and wide-covered Earth observation tasks. Hyperspectral image classification, which assigns each pixel of the hyperspectral image a semantic label by analyzing the corresponding spectral–spatial information, is the core technique of many Earth observation tasks [3]. However, due to the data redundancy and spectrum mixing, hyperspectral image classification is a hot yet challenging problem in remote sensing area [4]–[6].

During the past decade, motivated by its excellent learning ability, machine learning methods are extensively used for hyperspectral image classification [7]–[11]. In order to achieve acceptable performance, most machine learning-based methods work in a supervised manner [12]–[14], which takes some pixels with category information, i.e., labeled samples, of the hyperspectral image to build a training set, then trains a classification model on the training set, and, finally, classifies all the related unlabeled samples by the trained model. The most competitive supervised hyperspectral image classification methods are developed based on deep learning theory, which learns to extract representative and discriminative features by deep and hierarchical networks [15]–[18]. Although these machine learning-based methods make valuable explorations on hyperspectral image classification and significantly improve the performance of Earth observation tasks, they can reach satisfactory accuracy under severe conditions, i.e., when they are trained with sufficient training samples of the target data set, i.e., the data set of interest. However, since field investigation depends on related devices and land-cover experts, collecting sufficient labeled training samples is not practical in most remote sensing tasks. We believe that the knowledge of other data sets can reduce the dependence on sample quantity and assist the classification task on the target data set. Moreover, the cross-data set knowledge can also promote the training efficiency by the shared knowledge. Therefore, the exploitation of cross-data set knowledge can make the classification more practical and efficient.

The cross-data set knowledge is vital for hyperspectral image classification, but extracting and utilizing the cross-data set knowledge are not an easy task in the remote sensing area. Due to the environment change and equipment difference, domain shift is a very common issue, which troubles knowledge transferring and sharing in hyperspectral image classification. Therefore, some works utilize domain adaptation methods to promote the performance on the target data set

Manuscript received April 17, 2020; revised June 14, 2020 and July 9, 2020; accepted August 5, 2020. Date of publication August 18, 2020; date of current version April 22, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61801078, Grant 61671103, and Grant U1933104; in part by the China Postdoctoral Science Foundation under Grant 2018M630288; in part by the Liaoning Province Natural Science Foundation under Grant 20180520026; and in part by the Dalian Science and Technology Innovation Foundation under Grant 2018J12GX044. *(Corresponding author: Xiaorui Ma.)*

Xiaorui Ma, Xuerong Mou, and Hongyu Wang are with the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: maxr@dlut.edu.cn; mxr123@mail.dlut.edu.cn; whyu@dlut.edu.cn).

Jie Wang is with the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China, and also with the School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China (e-mail: wangjie@dlut.edu.cn).

Xiaokai Liu is with the School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China (e-mail: xkliu@dlmu.edu.cn).

Jie Geng is with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710068, China (e-mail: gengjie@nwpu.edu.cn).

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

Digital Object Identifier 10.1109/TGRS.2020.3015357

and uses the cross-data set knowledge as auxiliary information to effectively improve the related task on the target data set [19]–[22]. Some others also try to introduce source data sets to improve the classification performance using small training sets [23]–[25]. The aforementioned domain adaptation-based methods improve the classification accuracy on the target data set and designate the direction for the future development of the hyperspectral image classification. However, some of the aforementioned methods still require plenty of labeled samples on the target data set. Moreover, some domain adaptation-based methods only try to minimize domain discrepancy without considering the class distributions on the target domain, which drives the class boundaries crash into each other and, thus, reduces the effectiveness of the classifier from the source domain. In order to exploit and utilize the cross-data set knowledge, we propose a new hyperspectral image classification method that could drive the target domain to adapt to the source domain without any labeled samples in the target domain while keeping the boundaries of different classes.

In particular, this article proposes an adversarial domain adaptation method to exploit and utilize the cross-data set knowledge. Other than most existing methods, the proposed method gives a solution of cross-data set knowledge transferring when no labeled information of the target data set is available. Moreover, the proposed method is based on variational autoencoders and adversarial learning, which can align the source domain and the target domain while keeping the boundaries of different classes. The major contributions can be summarized as follows.

1) In order to learn the cross-data set knowledge, we develop a generator based on variational autoencoders, which learns spectral and spatial features of both domains. The generator can learn the distribution of the input and generate a more generalized representation from the learned distribution, which is more robust for the cross-data set classification problem.
2) In order to transfer the cross-data set knowledge, we design two objective functions for adversarial learning, which considers both global alignment and local alignment based on the classification error and the classification disagreement. The two alignment strategies help to align different domains while keeping the boundaries of different classes.
3) In order to achieve cross-data set classification, we train the whole network based on adversarial learning, which first initializes both the generator and the discriminator on the source domain and then adjusts the discriminator on both domains and fine-tunes the generator on the target domain in an adversarial manner by multiple times. The adversarial learning training process drives the target samples under the support of the source domain.

The rest of this article is organized as follows. A brief introduction of the related work is presented in Section II. The detailed implementation of the proposed method is given in Section III. The experiments with corresponding analysis are demonstrated in Section IV. The conclusions are summarized in Section V.

II. Related Work

In this section, we present a brief summary of the related methods in hyperspectral image classification, including deep network-based feature learning methods and domain adaptation based on classification methods.

A. Deep Learning

Inspired by its excellent ability to feature representation, deep learning has been widely used in hyperspectral image classification and improves the performance significantly. Different types of deep networks have been improved according to the specialty of the hyperspectral image, such as stacked autoencoders, convolution neural networks are used for spectral and spatial feature learning and information optimizing [15], [18], the generative adversarial network is utilized to generate fake samples and improve the generalization of the trained model [16], [26], and the recurrent neural network is employed for sequential analysis and classification of hyperspectral image [27], [28]. Some representative works show the effectiveness of deep learning in hyperspectral image classification. Chen *et al*. [18] combined metric learning with convolution neural network to alternately learn discriminative spectral–spatial features. Zhu *et al*. [16] took the advantage of the generative adversarial network to improve the generalization capability of the learned features, and Zhang *et al.* [28] proposed a novel local spatial sequential method to extract local and semantic information for hyperspectral image classification. Although deep learning has gained enough attention in hyperspectral image classification, it has an inevitable disadvantage, and the excellent learning ability is ensured by the tons of network parameters, which require lots of labeled samples to learn. However, due to the vast coverage of remote sensing images, field investigation is a time cost work, which makes collecting labeled training samples an extremely difficult task.

In order to improve the performance of deep learning-based methods using small training sets, the mainstream methods try to reorganize the training samples or borrow more information from other data sets. Li *et al*. [29] utilized pixel pairs as training samples to ensure a sufficient amount of training data and, hence, learn a large number of parameters, and Jiao *et al*. [30] used the parameters of ImageNet to initialize multiple parameters and reduce the difficulty of network training. Mei *et al*. [31] learned sensor-specific spatial–spectral features by sharing the parameters between the data from the same sensors. All these methods improve the classification accuracy under a harsh situation of small labeled training sets. In the case of a neural network with a large number of parameters, the efficiency is greatly improved. Inspired by the excellent learning ability of deep networks, this article proposes a feature learning method based on variational autoencoders for domain adaptation scenarios. Other than most existing methods, the proposed method learns the distribution of the input and generates more generalized representation without any labeled sample of the target data set.

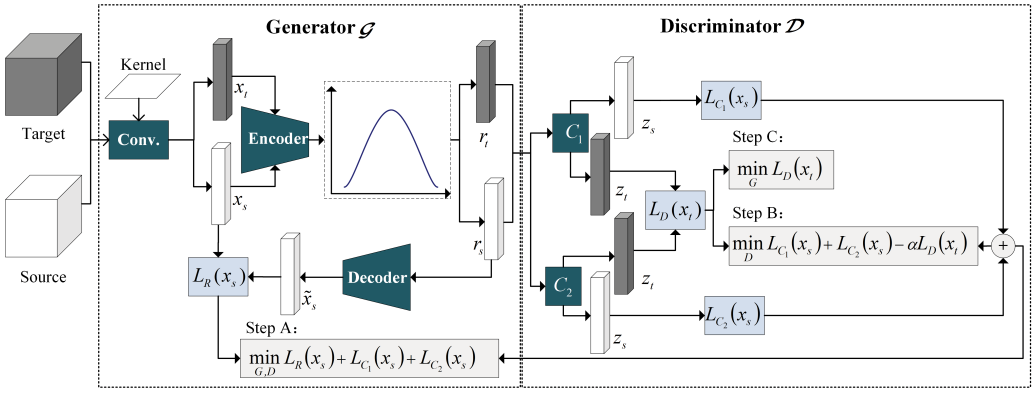


Fig. 1. Overview of the proposed adversarial domain adaptation method. There are two modules, i.e., a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$. $\mathcal{G}$ is a feature extractor based on variational autoencoders that learn the spectral–spatial features. $\mathcal{D}$ is built by multiple classifiers, which classifies the spectral–spatial features. Following the objective functions, three basic steps are utilized or repeated to train the whole deep network, Step A learns knowledge on the source domain, Step B Adjusts the discriminator on the target domain, and Step C Fine-tunes the generator on the target domain.

B. Domain Adaptation

Transfer learning methods are used to solve the problem of nonstationary data sets over time or space. Domain adaptation, which is an important member of the transfer learning family, attempts to adapt the model trained on the source domain to the target domain. Due to the capturing environment change and equipment difference, domain shift is common in remote sensing areas; hence. domain adaptation is necessary for cross-data set knowledge transferring and sharing in hyperspectral image classification [32].

During the past decade, many domain adaptation methods have been proposed for hyperspectral image classification. Some works attempt to transfer knowledge from different domains to improve the classification performance with limited labeled samples from the target domain [19]–[21], [33]. Qin *et al.* [34] presented a tensor alignment-based domain adaptation method for hyperspectral image classification, and Qin *et al.* [35] also introduced a novel heterogeneous domain adaptation method for hyperspectral image classification with a limited amount of labeled samples in both domains. Zhou *et al.* [36] proposed extreme learning machine-based heterogeneous domain adaptation algorithms for the classification of remote sensing images. The aforementioned methods rely on the labeled trained samples in the target domain, which reduces the practicability of hyperspectral image classification in remote sensing applications. Some other works utilize unsupervised domain adaptation methods to transfer knowledge to the target domain. Wei *et al.* [37] proposed an unsupervised domain adaptation method from both feature and classifier levels. Peng *et al.* [38] proposed a discriminative joint matching method to match source and target features in the space produced by kernel principal components. Gao *et al.* [39] developed an interesting work to align different domains in the tensor space and achieve excellent performance. However, in order to compromise to the strong constrain of global alignment of different domains, i.e., minimizing the domain discrepancy, some boundaries of different classes in the target domain may crash into each other, which incurs the performance degradation. In this article, the proposed method transforms target samples under the support of the source while keeping the boundaries of different classes; hence, the knowledge can be transferred and shared between data sets of different domains.

III. Methodology

In this article, we propose a domain adaptation method based on adversarial learning to exploit the cross-data set knowledge for hyperspectral image classification, which is shown in Fig. 1. The proposed method achieves adversarial learning by two modules: a generator based on variational autoencoders and a discriminator built by multiple classifiers. The two modules work in an adversarial manner to achieve domain aliment, hereby to realize the classification on the target domain in an unsupervised manner. First, the problem setting is given in Section III-A for the problem statement and symbol definition, and then, two major modules of the proposed method are depicted in Sections III-B and III-C, respectively. Finally, the whole framework of the proposed adversarial domain adaptation method is summarized in Section III-D.

A. Problem Setting

For a clear presentation, we first introduce the problem statement and explain the symbols used in the proposed method. We consider about two hyperspectral images $\boldsymbol{H}_s$ from the source domain and $\boldsymbol{H}_t$ from the target domain. We assume that the source domain and the target domain share the same land-cover types. Suppose that the source data set, i.e., the sample set from the source domain, is denoted as $\boldsymbol{X}_s = \{\boldsymbol{x}_s^{(i)}\}_{i=1}^{M}$, and the corresponding label set is denoted as $\boldsymbol{Y}_s = \{y_s^{(i)}\}_{i=1}^{M}$, where M is the number of samples in the source data set. The target data set, i.e., the sample set from the target domain, is represented by $\boldsymbol{X}_t = \{\boldsymbol{x}_t^{(j)}\}_{j=1}^{N}$, and the corresponding label set is represented by $\boldsymbol{Y}_t = \{y_t^{(j)}\}_{j=1}^{N}$, where N is the number of samples in the target data set. The label set of the target set is not available during the training process. Since the land-cover types are consistent, $y_s^{(i)}, y_t^{(j)} \in \{1, 2, \ldots, K\}$, where

K is the number of classes. Particularly, in order to learn both the spatial and spectral features, we utilize data cube including the spectral information of both the pixel under processing and the pixels from the neighboring area as the samples in both $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$. For implementation, each cube is cropped surrounding the pixel under processing from the original hyperspectral image, and the corresponding label is the label of the center pixel of the small cube.

Due to the capturing environment change and imaging sensor difference, even with the same land-cover types, the data distributions of $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ are different. Therefore, the domain shift is the major obstacle that causes the trained classification model failure on the target data set. The objective of this article is to align the target domain and the source domain to minimize the discrepancy between the two domains and make the classification model keep effective in the target domain. The proposed method only utilizes the label information from the source domain, i.e., $\boldsymbol{Y}_s = \{y_s^{(i)}\}_{i=1}^{M}$, and transfers the cross-data set knowledge from the source data set $\boldsymbol{X}_s$ to the target data set $\boldsymbol{X}_t$ without using any label of the target data set. Therefore, $\boldsymbol{X}_t$ acts as a testing set, whose label information is not used for training, i.e., $\boldsymbol{Y}_t$ is not available during the training process. As a result, the proposed method can realize classification on the target domain using only the cross-data set knowledge.

As shown in Fig. 1, the proposed method consists of two modules: a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$. $\mathcal{G}$ is based on improved variational autoencoders that learn the spectral–spatial features with the minimized classification disagreement. $\mathcal{D}$ is built by multiple classifiers, which classifies the spectral–spatial features with minimized classification error and maximized classification disagreement. The two modules work in an adversarial manner and try to drive the target samples under the source support by fine-tuning the feature embedding parameters and adjusting the classification parameters in turn. More detailed information on the generator and the discriminator will be explained in the following parts.

B. Generator: VAE-Based Feature Learning

In this section, we design a generator $\mathcal{G}$ based on variational autoencoders (VAE) to learn the spectral–spatial features and to drive the target domain under the support of the source domain in the feature space.

Intuitively, if we can learn the distribution of the training set and then sample from the learned distribution randomly to reconstruct a training set, we can represent the input samples more generally. VAE is designed based on this theory, learns the distribution of the input data set, and resamples from the learned distribution to build a more generalized training set. With VAE, we can extract more effective features for hyperspectral image classification. However, learning the distribution of the overall input data is difficult theoretically, and the resampling may cause mismatching between the input sample and the corresponding latent variable. Therefore, VAE tries to assume the conditional probability distribution of a latent variable that is specific to each input sample, which can be realized by minimizing the variational lower bound on the marginal likelihood function.

Mathematically, given a sample $\boldsymbol{x}$, VAE tries to find the conditional distribution of a latent variable $\boldsymbol{r}$ for the encoder, i.e., $p(\boldsymbol{r}|\boldsymbol{x})$, and utilize another distribution $q(\boldsymbol{r}|\boldsymbol{x})$ to approximate $p(\boldsymbol{r}|\boldsymbol{x})$. In practice, we assume that $q(\boldsymbol{r}|\boldsymbol{x})$ is a multivariate standard Gaussian distribution $N(\mu, \sigma)$. VAE utilizes an encoder to learn μ and σ of the Gaussian distribution and then resamples from $N(\mu, \sigma)$ to capture the latent variable $\boldsymbol{r}$ by a reparameterization strategy [40]. Finally, a decoder is used to reconstruct $\boldsymbol{x}$ from the latent variable, and the conditional distribution can be expressed as $p(\boldsymbol{x}|\boldsymbol{r})$. Moreover, in order to ensure that $q(\boldsymbol{r}|\boldsymbol{x})$ is similar to $p(\boldsymbol{r})$, which can be assumed as a standard normal distribution, the KL-divergence is utilized to minimize the difference between the two distributions. The objective function L_{R} of VAE is defined with probability function, which is

$$L_{\mathrm{R}} = \mathbb{E}_{q(\boldsymbol{r}|\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{r})] - \beta \mathbb{D}(q(\boldsymbol{r}|\boldsymbol{x})||p(\boldsymbol{r})) \tag{1}$$

where $\mathbb{E}$ is the arithmetic mean function and $\mathbb{D}$ is the KL-Divergence function. The first term is related to the reconstruction error, the second term is the KL-Divergence, and β is the importance weight. When the loss descends to an acceptable threshold or changes within a very small range, the hidden variable $\boldsymbol{r}^{(i)}$ is utilized as the feature of the input sample $\boldsymbol{x}^{(i)}$ and inputs into the following classification unit.

For hyperspectral image, we think that the spatial information is as much important as the spectral information, especially for the ones with high spatial resolution. In order to learn both the spectral information and spatial information, we equip VAE with one more convolutional layer at the beginning of the VAE network for spatial information learning and utilize the original VAE structures for spectral information learning. Mathematically, for each sample, i.e., data cube, we perform spatial filter with a 2-D Gaussian kernel and then input the spatial features into the following VAE structure to further learn the spectral features. We suppose that the embedding function over the whole feature learning modular is g, and then, $\boldsymbol{r}^{(i)} = g(\boldsymbol{x}^{(i)})$, which is the input feature of the next module.

C. Discriminator: Multiple-Classifier-Based Discriminator

More than enforcing global alignment between the source domain and the target domain, the proposed method tries to perform local alignment strategy, which drives each class in the target data set under the support of a corresponding class in the source data set, hereby keeps the performance of the classifiers trained on the source domain. The proposed method builds a discriminator based on multiple classifiers, minimizes the classification error to adjust the classifiers fit the source samples closely, and minimizes the classification disagreement to drive the target samples under the good support of the source domain. Before start, we should give the definition of classification disagreement of multiple classifiers, i.e., which can distinguish the target samples that are not well classified by the classifier trained on the source domain. We train multiple classifiers on the source domain and utilize

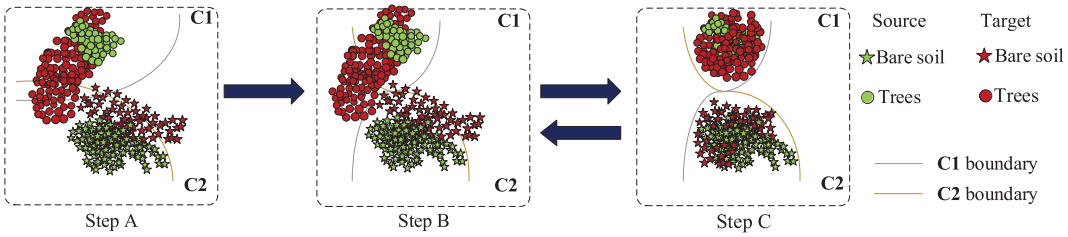


Fig. 2. Training procedure of the adversarial domain adaptation. Different colors mean different domains, and different shapes indicate different classes. Step A initializes the whole network by minimizing (4), Step B tries to push the classifier boundaries surrounding the source features tightly by minimizing (5), and Step C drives the target features under the support of the source domain by minimizing (6).

the disagreement of multiple classifiers to find out the target samples with poor support.

Suppose a classifier trained on the source data set, which takes the features of the source data set as input and learns to minimize the classification error on the source domain. With some classifiers, such as Softmax, we can get the classification probability of the sample belongs to each class from both classifiers, and the class with the highest probability should be the label of the input sample. Suppose that the embedding function of the classification is d, and then, the classification output is $\boldsymbol{z}_s^{(i)} = d(\boldsymbol{r}_s^{(i)})$. The classification loss function is defined as

$$L_C(\boldsymbol{X}_s) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} 1\left[y_s^{(i)} = k\right]\log\left(\boldsymbol{z}_s^{(i)}(k)\right) \tag{2}$$

where K is the number of classes, M indicates the number of source domain samples, y_s represents the true labels of sample $\boldsymbol{x}_s$, $\boldsymbol{z}_s^{(i)}(k)$ denotes the k^{th} element of $\boldsymbol{z}_s^{(i)}$, and $1[y_s^{(i)} = k]$ is the indicator function, which returns to 1 if the equation is true and 0 otherwise. It indicates the general classification error. All parameters can be learned by minimizing the classification loss function.

To distinguish the target samples without good source support, we first train two classifiers *i.e.*, C_1 and C_2, on the source domain by minimizing the classification loss function. Suppose that the embedding functions of C_1 and C_2 are d_1 and d_2, respectively. We then take the target features $\boldsymbol{r}_t^{(j)}$ as input and output the classification probability $d_1(\boldsymbol{r}_t^{(i)})$ and $d_2(\boldsymbol{r}_t^{(i)})$. We control the inconsistency of classification in the target domain using the output difference of the two classifiers. The objective function of classification disagreement is defined as follows:

$$L_{\mathcal{D}}(\boldsymbol{X}_t) = \frac{1}{N}\sum_{j=1}^{N}\left|\mathrm{d}_1\left(\boldsymbol{r}_t^{(j)}\right) - \mathrm{d}_2\left(\boldsymbol{r}_t^{(j)}\right)\right| \tag{3}$$

where N is the number of samples in the target data set.

From the loss function (refer to [41]), we can see that a larger disagreement loss guarantees the target samples are classified with low confidence, which indicates that they are far from the support of the source domain, i.e., close to the boundaries of the classifiers. By minimizing the classification disagreement, the target features can be classified with high confidence by the classifier trained on the source domain, which moves the target features under the support of the source domain.

D. Adversarial Domain Adaptation

The detailed framework of adversarial domain adaptation is shown in Fig. 2 and described in Algorithm 1. There are three major operations to perform the proposed adversarial domain adaptation method: first, learning knowledge on the source domain, which trains the generator and the discriminator only on the source data set to get the best performance on the source domain; then, adjusting the discriminator on both data set, which adjusts the classifiers closely fit the source data set while fixing the features; and finally, fine-tuning the generator on the target data set, which moves the target features under the support of the source features while fixing all classifiers of the discriminator. The last two steps work in an adversarial way to promote each other and align the two domains while keeping the effectiveness of the classifiers trained on the source data set.

Algorithm 1 Training Procedure

Require: The labeled source sample set $\{\boldsymbol{X}_s, \boldsymbol{Y}_s\}$, and the unlabeled target sample set $\boldsymbol{X}_t$.

1: Training the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$ with $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ on $\boldsymbol{X}_s$ using Eq. (4).
2: **repeat**
3: Fixing $\mathcal{G}$, adjusting the parameters of $\mathcal{D}$ on both $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ using Eq. (5).
4: Fixing $\mathcal{D}$, fine-tuning the parameters of $\mathcal{G}$ on $\boldsymbol{X}_t$ using Eq. (6).
5: **until** convergence

Ensure: Optimal parameters of $\mathcal{G}$ and $\mathcal{D}$

1) Step A. Learning Knowledge on the Source Domain: This step trains the whole network on the source data set, i.e., initializes both the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$ on the source data set. $\mathcal{G}$ is a feature extractor, which takes each sample $\boldsymbol{x}_s$ from the source data set as input and learns the spectral–spatial features $\boldsymbol{r}_s$, and $\mathcal{D}$ is consisted of two classifiers C_1 and C_2, which classifies the learned features. All parameters are learned by minimizing the reconstruction error and the classification error in the source domain. The objective function is defined as follows:

$$\min_{\mathcal{G},\mathcal{D}} L_R(\boldsymbol{X}_s) + L_{C_1}(\boldsymbol{X}_s) + L_{C_2}(\boldsymbol{X}_s) \tag{4}$$

where the subscripts under min means that we adjust all parameters in both $\mathcal{G}$ and $\mathcal{D}$ to achieve the training of the

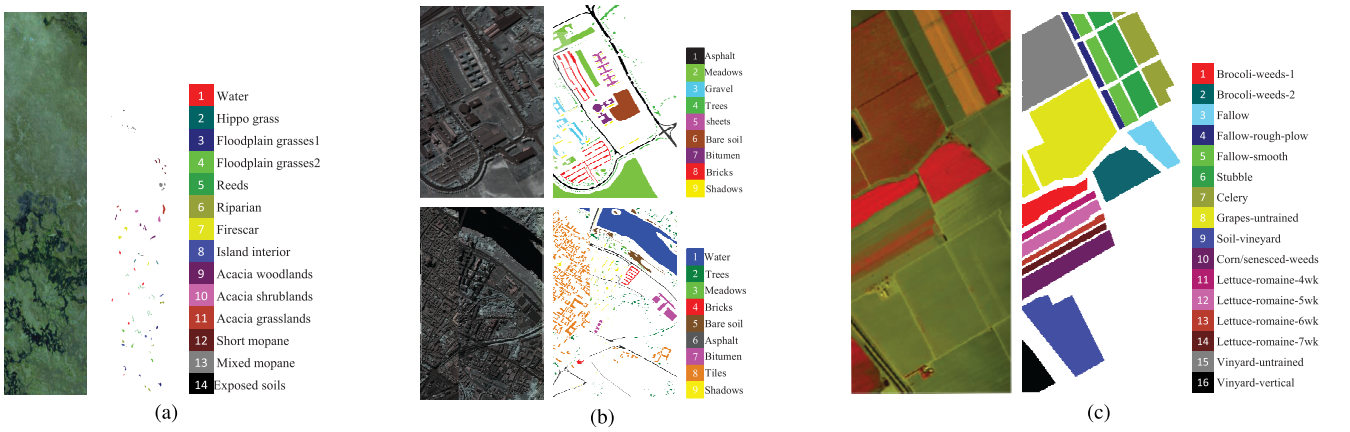


Fig. 3. All hyperspectral scenes used in this article, including the pseudocolor images and the corresponding ground truths with color indexes. (a) Botswana scene. (b) Pavia Scenes. (c) Salinas scene.

whole framework. This step makes the classifiers roughly fit the features of the source data set.

2) Step B. Adjusting the Discriminator on Both Domains: This step and the following Step C work in an adversarial manner to align the source domain and the target domain. While keeping the parameters of $\mathcal{G}$ fixed, this step adjusts the parameters of $\mathcal{D}$ to minimize the classification error on the source data set and maximize the classification disagreement on the target data set. The objective function is defined as follows:

$$\min_{\mathcal{D}} L_{C_1}(\boldsymbol{X}_s) + L_{C_2}(\boldsymbol{X}_s) - \alpha L_{\mathcal{D}}(\boldsymbol{X}_t) \tag{5}$$

where α is the importance weight, and the subscript under min means that it adjusts only the parameters in $\mathcal{D}$, i.e., changes the parameters of the feature extractor. This step is executed in both domains. On the source domain, it drives the two classifiers of the discriminator closely fit to the source features, i.e., moves the classifier boundaries close to the source class. On the target domain, it adjusts the discriminator to maximize the classification disagreement of the target features and, therefore, resists the following step to move the target domain under the support of the source domain.

3) Step C. Fine-Tuning the Generator on the Target Domain: This step works in an adversarial manner with Step B to move the target features under the support of the source features. While keeping the parameters of $\mathcal{D}$ fixed, this step fine-tunes the parameters of $\mathcal{G}$ to minimize the classification disagreement. The objective function is given as follows:

$$\min_{\mathcal{G}} L_{\mathcal{D}}(\boldsymbol{X}_t) \tag{6}$$

where the subscript under min means that it fine-tunes only the parameters in the $\mathcal{G}$, i.e., changes the parameters of the two classifiers. This step is only performed on the target domain, which fine-tunes the generator to resist the previous step. In the end, it makes the generator producing target features with a small intraclass difference and close to the source features, hence moving the target domain under the support of the source domain.

During training, we first execute Step A to initialize both the generator and the discriminator on the source domain for once and then carry out Step B on both domains and Step C on the target domain to further adjust the network in an adversarial manner for multiple times. Especially, sometimes, Step C is performed more than one time to minimize the classification disagreement. Adaptive moment estimation (usually know as Adam) [42] is used as optimization method to achieve the minimization of the loss functions.

IV. Experiments and Analysis

In order to evaluate the performance of the proposed adversarial domain adaptation method (ADA-Net), we compare the classification results of ADA-Net with five other related methods along with corresponding discussion. We also give a detailed analysis of all parameters related to the proposed ADA-Net. All the experiments are implemented on a desk computer with Intel Core i7 4.0-GHz CPU, GeForce GTX 1080Ti GPU, and 32-GB memory. Moreover, PyTorch with Python 3.6 is used to implement the deep network and improve programming efficiency.[1]

A. Experimental Setup and Data Set

Three scenes are used for experiments, i.e., Botswana scene, Salinas scene, and the Pavia scene with two data sets: the University of Pavia data set, and the Pavia Center data set. All four data sets and the related experimental setup are shown in Fig. 3 and Table I.

1) The data set of the Botswana scene used in this article was collected by a Hyperion imager on the NASA EO-1 satellite over the Okavango Delta, Botswana, 2001. After removing bands of water absorption and low SNR, 145 bands left. Moreover, for better display, we also cut out a small region without any labeled samples from the top left 1111 × 256 pixels. The spatial and spectral resolutions are 30 m and 10 nm, respectively; 3248 samples from 14 classes of land-cover are labeled. The pseudocolor image and the available ground-truth map with corresponding color index are

[1]The PyTorch toolbox is available at https://pytorch.org/

TABLE I

DETAILED INFORMATION OF ALL DATA SETS, INCLUDING SENSORS, WIDTH × HEIGHT × BAND NUMBER, SPATIAL AND SPECTRAL RESOLUTIONS, CLASS NUMBER OF THE INTERESTED TASK (N_c), THE SETUP OF THE SOURCE, AND THE TARGET DOMAIN ($\boldsymbol{H}_s$,$\boldsymbol{H}_t$)

Dataset	Sensor	Size	Resolution	N_c	$\boldsymbol{H}_s$	$\boldsymbol{H}_t$
Botswana	Hyperion	1111×256×145	30m, 10nm	12	1111×(0-128)×145	1111×(129-256)×145
Pavia Scenes	ROSIS	1096×715×102, 610×340×102	1.3m, 4nm	7	University of Pavia	Pavia Center
Salinas	AVIRIS	512×217×204	3.7m, 10nm	16	512×(0-25,101-150)×204	512×(26-100,151-217)×204

shown in Fig. 3(a). During experiments, we cut the original image into two subsets: the left half with 1150 known samples serves as the source data set, and the right half with 1649 known samples serves as the target data set. The common 12 land-cover types are used as the task of interest.

2) The data sets of the Pavia scene used in this article were collected by the Reflective Optics System Imaging Spectrometer over Pavia city, Northern Italy, 2002. Two data sets are collected over the Pavia scene, the University of Pavia and Pavia center, whose spatial and spectral resolutions are 1.3 m and 4 nm, respectively. After removing black regions, the spatial size are 1096×715 and 610×340, respectively. For experiments, 102 spectral bands are selected for both data sets. There are nine cover types in the reference data set. The pseudocolor image and the ground-truth map are shown in Fig. 3(b). During experiment, we utilize the University of Pavia with 39 332 known samples as the source data set and use another one with 39 355 known samples as the target data set. The common seven land-cover types are used as the task of interest.

3) Data set of the Salinas scene was captured by the Airborne Visible/Infrared Imaging Spectrometer over Salinas Valley, CA, USA, in 1998. This data set contains 512×217 pixels with 3.7-m spatial resolution and 224 spectral bands with 10-nm spectral resolution. Before experiment, 20 bands are abandoned due to water absorption. Except unknown samples, 16 classes are labeled in the available ground truth. The pseudocolor image and the ground-truth map are shown in Fig. 3(c). Since there is no other data set with the same land-cover types as Salinas scene, we cut the original image into four subsets, select two of them with 19 490 known samples as the source data set, and utilize the left with 34 639 known samples as the target data set.

B. Parameters Analysis

In order to give a comprehensive study of the proposed ADA-Net, we analyze all key parameters, including the network parameters and the training parameters. The former ones are related to the network structure, such as layer number, unit number, and kernel size. The latter ones affect the training process, such as the impact of training size and the number of cycles in Step C.

1) Network Parameters: We analyze all network parameters, including the number of layers, the number of units in each fully-connected layer, and the size of kernels in the only convolutional layer. All the abovementioned network parameters determine the configuration of the network. Particularly, the number of kernels is fixed to 1, i.e., we use the same kernel for all bands to extract spatial features. Three data sets are utilized to evaluate the performance under different network configurations; the analysis results evaluated by OA(%), AA(%), and $\kappa(\times 100)$ are summarized in Table II.

TABLE II

NETWORK PARAMETERS ANALYSIS [EVALUATED BY OA(%), AA(%), AND $\kappa(\times 100)$] OF THE PROPOSED ADA-NET USING ALL TARGET DATA SETS

Dataset	No.	Network Architecture	OA(%)	AA(%)	$\kappa(\times 100)$
Botswana	1	145(13×13)-100-40	85.20	81.33	80.16
	2	**145(13×13)-80-30**	88.96	87.93	84.83
	3	145(13×13)-80-20	86.86	85.31	82.38
Pavia Center	4	102(17×17)-70	81.40	77.88	76.51
	5	**102(17×17)-70-30**	87.68	83.73	85.26
	6	102(17×17)-70-50-30	84.67	84.50	81.63
Salinas	7	204(15×15)-130-50	84.67	80.51	80.33
	8	**204(17×17)-130-50**	86.17	87.60	81.27
	9	204(19×19)-130-50	83.46	79.35	81.20

During experiments, we keep other parameters fixed to verify the influence of a certain network parameter. We use the Botswana scene to validate the effect of unit number in Lines 1–3 in Table II. From the table, we can see that overmuch units will bring redundant information and increase computing costs, but insufficient units will be lost data information. The best unit number setup is shown in line 2. Moreover, we use the Pavia Center data set to test the influence of the number of layers in Lines 4–6. We can conclude that lacking layers may lead to incomplete information, while overmuch layers may cause too many parameters to learn. The best layer setup is shown in line 5. Finally, for the kernel size of the only convolutional layer, the data collected by different sensors show different kernel sizes. We think that the size of the convolutional kernel is closely related to the spatial resolution of hyperspectral images. Due to the difference in spatial resolution, the Salinas scene with spatial resolution 3.7 m has larger kernel size than the Botswana data set with spatial resolution 30 m.

2) Training Parameters: The training parameters are closely related to the training and optimization process, including the impact of training size and the number of cycles in Step C. All target data sets are used in the training parameters analysis, and the curves of the relation between these factors and the classification performance [represented by OA(%)] are shown in Fig. 4.

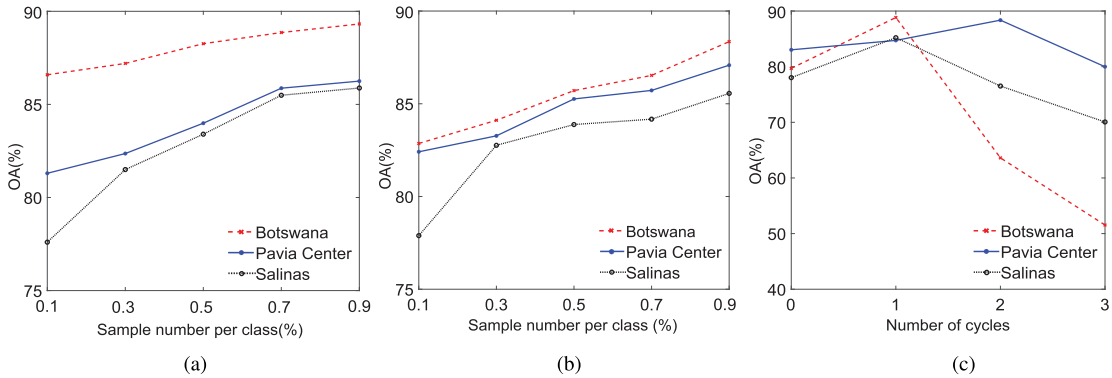


Fig. 4. Impact analysis of several parameters for all three target data sets of the proposed ADA-Net. (a) Training size of the source data set. (b) Training size of the target data set. (c) Number of cycles in Step C .

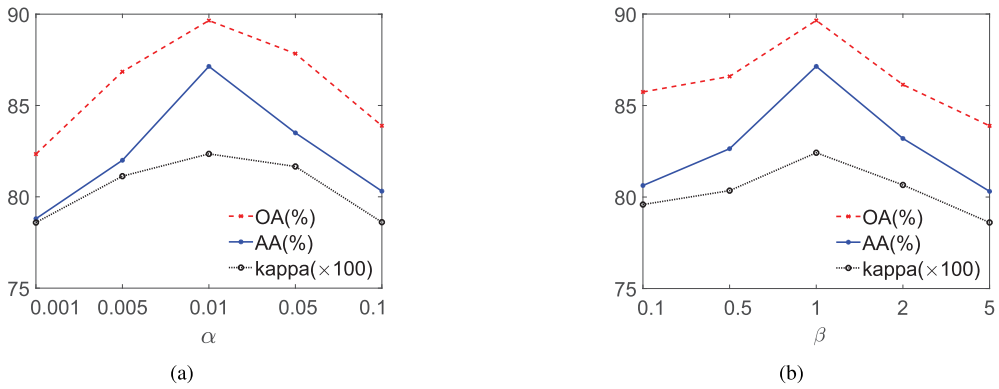


Fig. 5. Impact analysis of α and β for the Botswana data set of the proposed ADA-Net. (a) Parameter α. (b) Parameter β.

We analyze the sizes of training sets using all three target data sets, including the labeled training set of the source domain and the unlabeled training set of the target domain. The training set of the source domain changes from 10% to 90% of all labeled samples. As can be seen in Fig. 4(a), ADA-Net produces better as the number of source domain samples increases, and the performance tends to be stable when the training size is big enough, which indicates that the more the training samples from the source domain, the more the accurate cross-data set knowledge are learned. In Fig. 4(b), we present the relation between the number of unlabeled samples and the classification performance. Similarly, we can see that the more the training samples in the target domain, the higher the classification accuracy will be produced.

We also analyze the effect of the number of cycles in Step C and present the result in Fig. 4(c). Due to the nature of the data set itself, the numbers of cycles giving the best performance are slightly different. When the number of cycles is set to 0, the classification performance is close to the domain adaptation method using only global alignment. At the same time, we can find that the best cycle number is close to 1, i.e., 1 for both Botswana data set and Salinas data set, and 2 for Pavia Center. We think the reason is that when the number of cycles is too big, the output of the two classifiers will be too small so that the whole target data set will be driven into the same category.

We also take the Botswana data set as an example to analyze α and β and present the result in Fig. 5. α reflects the ability of local alignment in Step B. If it is difficult to achieve local alignment, and α should be larger and conversely smaller. In the method that we proposed, the setting parameters of α are 0.01 for the Botswana scene. β reflects that the decoder can be robust to the noise. It depends on the reality of the different data sets. The setting parameters of β are 1 for the Botswana scene.

C. Performance Analysis

We evaluate the performance of the proposed ADA-Net using comparison experiments. First, since the proposed method is a cross-data set classification method without any labeled sample on the target data set, we compare it with an unsupervised classification method based on a deep network (Un-Net). Un-Net serves as a baseline, and it utilizes stacked autoencoder to learn features and a clustering method to classify [43]. Second, to test the effectiveness of domain adaptation, we trained a network, namely, D-Net on the source data set, and directly transfer to the target data set. D-Net shares the same feature learning and classification structure as the proposed ADA-Net. Moreover, we compare the discriminative transfer joint matching (DTJM) method [38], and the method is based on the kernel principal component analysis method to match the distribution of the source domain and target domain,

TABLE III

CLASSIFICATION PERFORMANCE EVALUATED BY OA(%), AA(%), AND $\kappa(\times 100)$ OF DIFFERENT METHODS FOR ALL TARGET DATA SETS

Dataset	Measurement	Un-Net	Sup-Net	DTJM	FT-Net	D-Net	ADA-Net
Botswana	OA(%)	23.69±5.96	75.19±3.93	78.32±1.63	85.32±4.37	79.66±5.33	**89.65±2.20**
	AA(%)	20.94±3.92	75.09±4.21	76.91±2.17	83.11±3.28	74.88±3.21	**87.14±3.82**
	$\kappa(\times 100)$	21.01±7.91	72.36±3.15	75.06±5.33	83.17±5.23	75.71±7.25	**82.35±4.78**
Pavia Center	OA(%)	39.68±5.09	76.67±10.36	80.72±1.06	85.70±7.59	82.58±4.72	**88.25±3.01**
	AA(%)	41.31±6.58	77.36±5.41	75.94±2.14	86.92±6.67	81.87±5.69	**87.93±9.25**
	$\kappa(\times 100)$	38.72±7.06	72.16±6.28	77.97±4.61	84.13±8.25	76.22±4.15	**84.92±3.07**
Salinas	OA(%)	26.33±4.21	73.29±6.59	82.08±1.82	79.46±4.97	78.91±4.09	**86.55±2.44**
	AA(%)	20.99±4.69	69.83±5.36	77.35±3.01	80.71±5.16	80.09±5.28	**87.30±4.97**
	$\kappa(\times 100)$	22.36±6.33	68.78±5.86	72.78±6.69	75.73±5.32	75.59±3.45	**81.96±5.09**

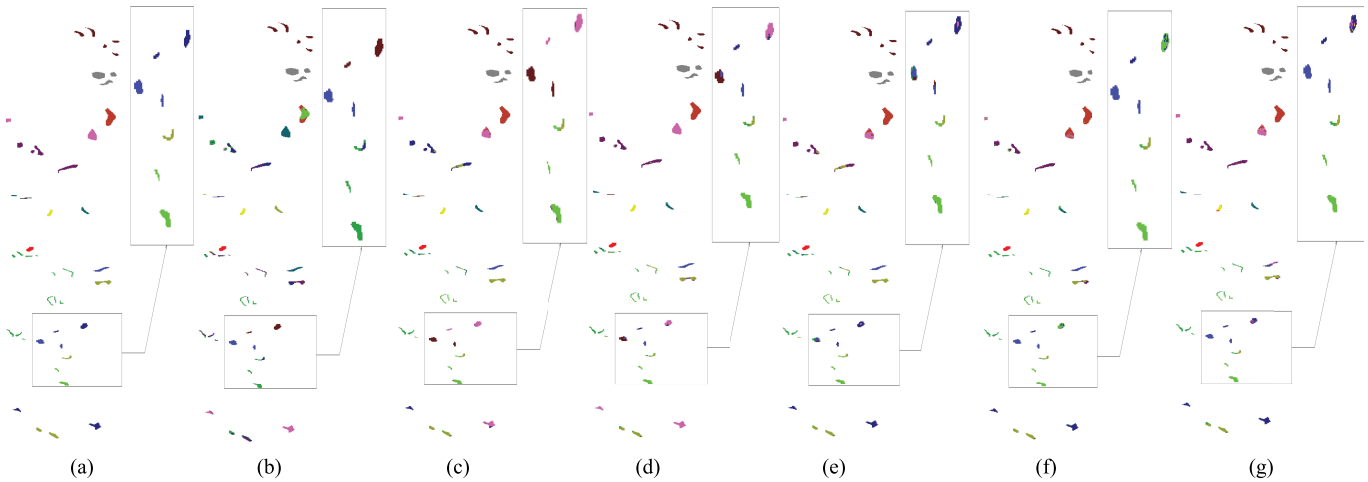


Fig. 6. Classification maps of the Botswana scene produced by different methods. (a) Ground truth. (b) Un-Net. (c) Sup-Net. (d) DTJM. (e) FT-Net. (f) D-Net. (g) ADA-Net.

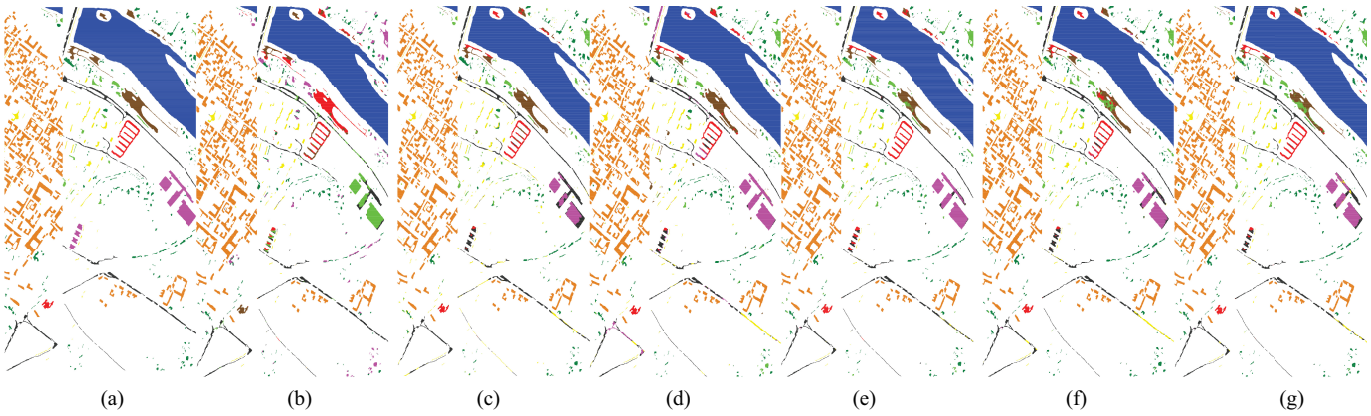


Fig. 7. Classification maps of the Pavia Center scene produced by different methods. (a) Ground truth. (b) Un-Net. (c) Sup-Net. (d) DTJM. (e) FT-Net. (f) D-Net. (g) ADA-Net.

which is a global alignment adaptive method, and serves as the state-of-the-art method. Furthermore, we try to challenge some supervised methods under a few-shot setting, a supervised classification method trained on the target data set, Sup-Net [44], and a supervised domain adaptation method based on fine-tuning, FT-Net [45]. FT-Net fine-tunes the trained network on the target domain.

For both the proposed ADA-Net and all the comparing methods, we use all labeled samples of the source domain and all unlabeled samples in the target domain for network training or clustering. We set α and β according to the parameter analysis curves, for the Botswana data set: 0.01 and 1, for Pavia scene: 0.00001 and 0.01, and for Salinas data set: 0.00001 and 0.001. All the corresponding OA, AA, and κ are recorded in Table III, and classification maps of all methods are shown in Figs. 6–8. From the figures, we conclude that our ADA-Net can produce the best classification performance for target domain data under unsupervised conditions. For the

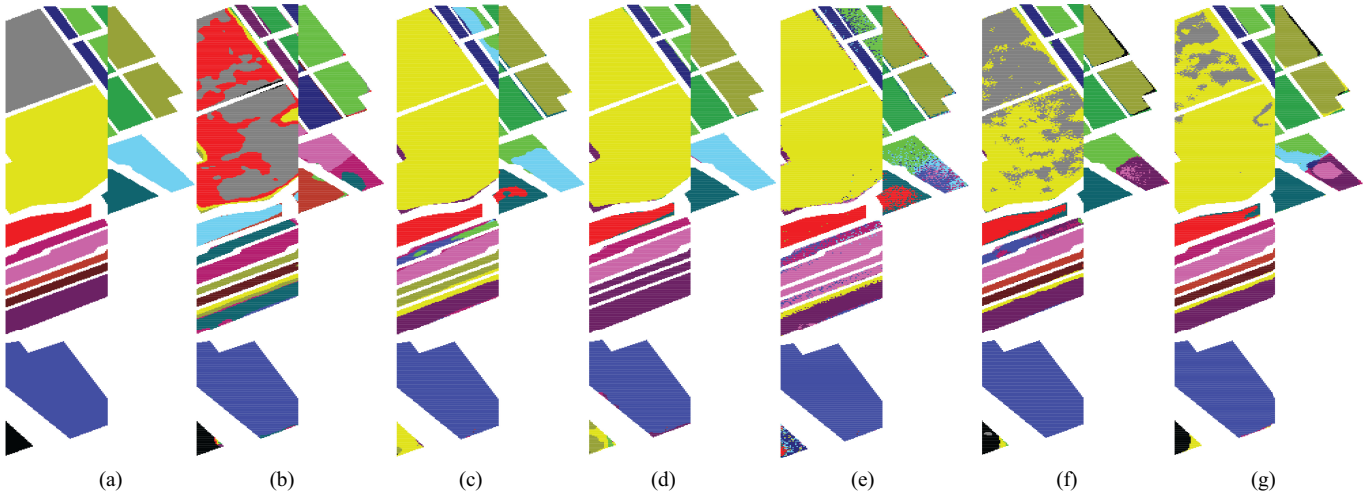


Fig. 8. Classification maps of the Salinas scene produced by different methods. (a) Ground truth. (b) Un-Net. (c) Sup-Net. (d) DTJM. (e) FT-Net. (f) D-Net. (g) ADA-Net.

TABLE IV
TRAINING TIME OF DIFFERENT METHODS FOR THE BOTSWANA DATA SET

Method	Un-Net	Sup-Net	DTJM	FT-Net	D-Net	ADA-Net
Time(s)	-	25	34	42	31	58

Botswana data set, FT-Net can produce the results closest to the method that we proposed. After our analysis, we found that the differences between the data set class are very small and, therefore, relatively easy for classification tasks, and a small number of training samples could achieve a good classification performance. The result of FT-Net is better than that of D-Net because there is one sample of the target domain to fine-tune the network.

Furthermore, we give an estimation of the execution time. For the Botswana data set, the time cost of all methods is shown in Table IV. Since the proposed network consists of only one convolutional layer and multiple fully connected layers, it contains far fewer parameters than many convolutional networks and, thus, can be easily trained with a relatively small training set. However, since the proposed method considers the adversarial network, the training time is a little longer than the methods that are only trained on the target domain. From the table, we can see that the training time of ADA-Net is about 58 s. About testing time, except UN-Net, all methods have almost the same testing time, about 0.04 s. UN-Net, which clusters all testing samples to classify, has the longest testing time, about 100 s. Therefore, the execution time of the proposed ADA-Net is acceptable for practical applications.

V. CONCLUSION

In order to exploit the cross-data set knowledge, this article proposes a domain adaptation method for hyperspectral image classification. The proposed method is based on adversarial learning, a VAE-based generator learns features to minimize the classification error on the source data set and maximize the classification disagreement on the target data set, and a multiclassifier-based discriminator adjusts classifiers to minimize the classification disagreement on the target data set. As a result, the proposed method can minimize the discrepancy of different domains while keeping the boundaries of different classes. Experimental results verify the effectiveness of the proposed ADA-Net on dealing with domain shift issue.

However, the proposed ADA-Net can only transfer between different data sets with the same land-cover types. We will study the proposed method more to deal with domain adaptation with different tasks as our further work.

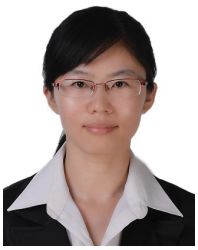
ACKNOWLEDGMENT

The authors would like to thank all researchers and organizations for kindly providing the experimental data set and the corresponding ground truths.

REFERENCES

[1] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, Jan. 2014.
[2] P. Ghamisi *et al.*, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.
[3] A. F. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Science*, vol. 228, no. 4704, pp. 1147–1153, Jul. 1985.
[4] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral–spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.
[5] Y. Gu, J. Chanussot, X. Jia, and J. A. Benediktsson, "Multiple kernel learning for hyperspectral image classification: A review," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6547–6565, Nov. 2017.
[6] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.
[7] J. C. Harsanyi and C.-I. Chang, "Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 4, pp. 779–785, Jul. 1994.
[8] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, "PCA-based edge-preserving features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7140–7151, Dec. 2017.
[9] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 99, pp. 1–5, Feb. 2018.

[10] L. Mou, P. Ghamisi, and X. Xiang Zhu, "Unsupervised spectral–spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.
[11] Q. Gao, S. Lim, and X. Jia, "Spectral–spatial hyperspectral image classification using a multiscale conservative smoothing scheme and adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7718–7730, Oct. 2019.
[12] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.
[13] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.
[14] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016.
[15] 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.
[16] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.
[17] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 120–147, Nov. 2018.
[18] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.
[19] X. Zhou and S. Prasad, "Deep feature alignment neural networks for domain adaptation of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5863–5872, Oct. 2018.
[20] J. Shen, X. Cao, Y. Li, and D. Xu, "Feature adaptation and augmentation for cross-scene hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 622–626, Apr. 2018.
[21] Z. Wang, B. Du, Q. Shi, and W. Tu, "Domain adaptation with discriminative distribution and manifold embedding for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 7, pp. 1155–1159, Jul. 2019.
[22] T. Liu, X. Zhang, and Y. Gu, "Unsupervised cross-temporal classification of hyperspectral images with multiple geodesic flow kernel learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9688–9701, Dec. 2019.
[23] C. Persello and L. Bruzzone, "Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2615–2626, May 2016.
[24] J. Lin, L. Zhao, S. Li, R. Ward, and Z. J. Wang, "Active-learning-incorporated deep transfer learning for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4048–4062, Nov. 2018.
[25] L. Windrim, A. Melkumyan, R. J. Murphy, A. Chlingaryan, and R. Ramakrishnan, "Pretraining for hyperspectral convolutional neural network classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2798–2810, May 2018.
[26] J. Feng, H. Yu, L. Wang, X. Cao, X. Zhang, and L. Jiao, "Classification of hyperspectral images based on multiclass spatial–spectral generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5329–5343, Aug. 2019.
[27] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.
[28] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, and H. Zhou, "Spatial sequential recurrent neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4141–4155, Nov. 2018.
[29] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.
[30] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, "Deep fully convolutional network-based spatial ddistribution prediction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5585–5599, Oct. 2017.
[31] S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, "Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4520–4533, Aug. 2017.
[32] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.
[33] X. Li, L. Zhang, B. Du, and L. Zhang, "On gleaning knowledge from cross domains by sparse subspace correlation analysis for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3204–3220, Jun. 2019.
[34] Y. Qin, L. Bruzzone, and B. Li, "Tensor alignment based domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9290–9307, Nov. 2019.
[35] Y. Qin, L. Bruzzone, B. Li, and Y. Ye, "Cross-domain collaborative learning via cluster canonical correlation analysis and random walker for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3952–3966, Jun. 2019.
[36] L. Zhou and L. Ma, "Extreme learning machine-based heterogeneous domain adaptation for classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1781–1785, Nov. 2019.
[37] W. Wei, W. Li, L. Zhang, C. Wang, P. Zhang, and Y. Zhang, "Robust hyperspectral image domain adaptation with noisy labels," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 7, pp. 1135–1139, Jul. 2019.
[38] J. Peng, W. Sun, L. Ma, and Q. Du, "Discriminative transfer joint matching for domain adaptation in hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 972–976, Jun. 2019.
[39] G. Gao and Y. Gu, "Tensorized principal component alignment: A unified framework for multimodal high-resolution images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 46–61, Jan. 2019.
[40] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," May 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114
[41] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[43] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.
[44] X. Ma, H. Wang, and J. Geng, "Spectral–spatial classification of hyperspectral image based on deep auto-encoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4073–4085, Sep. 2016.
[45] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Dec. 2014, pp. 3320–3328.

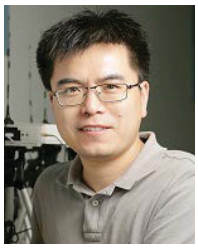


Xiaorui Ma (Member, IEEE) received the B.S. degree in applied mathematics from Lanzhou University, Lanzhou, China, in 2008, and the Ph.D. degree in communication and information system from the Dalian University of Technology, Dalian, China, in 2017.

She is a Lecturer with the Dalian University of Technology. Her research interests include processing and analysis of remote sensing images, especially hyperspectral image classification and synthetic aperture radar image classification.

Xuerong Mou received the B.S. degree in electronic and information engineering from Dalian Maritime University, Dalian, China, in 2018. She is pursuing the M.E. degree with the School of Information and Communication Engineering, Dalian University of Technology, Dalian.

Her research interests include hyperspectral image classification and target detection, remote sensing image analysis and interpretation, and machine learning.



Jie Wang (Senior Member, IEEE) received the B.S. degree from the Dalian University of Technology, Dalian, China, in 2003, the M.S. degree from Beihang University, Beijing, China, in 2006, and the Ph.D. degree from the Dalian University of Technology, in 2011, all in electronic engineering.

He is a Full Professor with Dalian Maritime University, Dalian. His research interests include wireless localization and tracking, radio tomography, wireless sensing, cognitive radio networks, and machine learning.

Xiaokai Liu received the B.S. degree in applied mathematics from Lanzhou University, Lanzhou, China, in 2008, and the Ph.D. degree in communication and information system from the Dalian University of Technology, Dalian, China, in 2017.

She is a Lecturer with Dalian Maritime University, Dalian. Her research interests include processing and analysis of remote sensing images, especially hyperspectral image classification and synthetic aperture radar image classification.

Jie Geng (Member, IEEE) received the B.S. and Ph.D. degrees in electronic and information engineering from the Dalian University of Technology, Dalian, China, in 2013 and 2018, respectively.

He is an Assistant Professor with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. His research interests include processing and analysis of remote sensing images, especially hyperspectral image classification and synthetic aperture radar image classification.

Hongyu Wang (Member, IEEE) received the B.S. degree in electronic engineering from the Jilin University of Technology, Changchun, China, in 1990, the M.S. degree in electronic engineering from the Graduate School of Chinese Academy of Sciences, Changchun, in 1993, and the Ph.D. degree in precision instrument and optoelectronics engineering from Tianjin University, Tianjin, China, in 1997.

He is a Professor with the Dalian University of Technology, Dalian, China. His research interests include image processing, image analysis, and remote sensing image classification.