Augmented Associative Learning-Based Domain Adaptation for Classification of Hyperspectral Remote Sensing Images

Min Chen, *Student Member, IEEE*, Li Ma, *Member, IEEE*, Wenjin Wang, *Student Member, IEEE*, and Qian Du, *Fellow, IEEE*

***Abstract*—Associative learning-based domain adaptation approach is investigated for the classification of hyperspectral remote sensing images in this article. It employs the criterion of cycle consistency to achieve features that are both domain-invariant and discriminative. Two cross-domain similarity matrices based on network-generated features and probability predictions are introduced in the two-step transition procedure. The associative learning with feature and prediction-based similarity metrics is referred to as augmented associative learning (AAL). The AAL-based domain adaptation network does not require target labeled information and can achieve unsupervised classification of the target image. The experimental results using Hyperion and AVIRIS hyperspectral data demonstrated the efficiency of the proposed approach.**

***Index Terms*—Associative learning, classification, domain adaptation, hyperspectral remote sensing.**

I. Introduction

DURING the past years, a large number of research efforts have been spent on remote sensing applications, such as classification [1]–[4], feature extraction [5], target detection [6], etc. Due to the high spectral resolution, hyperspectral images (HSIs) are superior to other remote sensing images in classification-based tasks when spectral features are similar. Thanks to the developed hyperspectral sensor, it is easy to collect a large number of HSIs. However, the acquisition of labeled samples is costly, labor-consuming, or even infeasible. Therefore, it motivated us to train an effective classifier on available labeled HSI, which may work well on another image with few or without labeled data. However, spectral features may vary significantly between multitemporal images or spatially disjoint images because of changed soil moisture, vegetation composition, topography, and sun angle [7], [8]. Inevitably, it will cause a poor classification performance when a classifier trained on other images is directly applied to a new image. This problem can be tackled by domain adaptation (DA) [9], [10] which attempts to train a model by transferring knowledge from a label-rich source domain to a target domain with few or no labels [11].

DA methods can be classified into semisupervised and unsupervised forms according to the availability of target labeled data [12]. Suppose there exist abundant source labeled data. When a few target labeled data are available, the DA method is identified as a semisupervised method, and when target labels are not accessible, it is an unsupervised approach. For semisupervised DA methods, the common DA strategies include utilizing both source labels and target labels to train a classifier [13], pretraining a network with source labeled data and fine-tuning it with target labeled data [14], learning a common feature space by using the corresponding data pairs [15], and so on. Unsupervised DA has no access to the target labels which is a more challenging task. In this article, we focus on the unsupervised domain adaptation.

Unsupervised DA methods can be approximately divided into three categories: divergence-based adaptation, adversarial-based adaptation, and reconstruction-based adaptation. Reconstruction-based adaptation methods are achieved by constraining that one domain can be well reconstructed by the other domain. Wang *et al.* proposed a classwise reconstruction-based adaptation method, which fully exploited intraclass dependence and interclass independence [16]. Jhuo *et al.* presented a low-rank reconstruction method to reduce the domain disparity [17]. Divergence-based adaptation relied on divergence measures between the source and target domains. Maximum mean discrepancy (MMD) [18] is the most popular DA strategy. Gopalan *et al.* first used MMD to reduce the distribution mismatch between two hidden layer representations in the context of neural network [19]. Tzeng *et al.* proposed a new convolutional neural network, which used an adaptation layer along with a domain confusion loss based on MMD to learn domain-invariant representations [20]. In [21], three adaptation layers based on multiple kernel variant of MMD (MK-MMD) were utilized to learn task-specific features and enhance adaptation effectiveness in a deep adaptation network (DAN). Similarly, correlation alignment (CORAL) [22] can match the

Manuscript received June 11, 2020; revised September 6, 2020; accepted September 28, 2020. Date of publication October 13, 2020; date of current version October 23, 2020. This work was supported by National Natural Science Foundations of China under Grant 61771437, Grant 61102104, and Grant 91442201. *(Corresponding author: Li Ma.)*

Min Chen, Li Ma, and Wenjin Wang are with the School of Mechanical Engineering and Electronic Information, China University of Geosciences,Wuhan, Hubei 430074, China (e-mail: suprecm7@cug.edu.cn; mali@cug.edu.cn; crazyjin1996@cug.edu.cn).

Qian Du is with the Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS 39762 USA (e-mail: du@ece.msstate.edu).

Digital Object Identifier 10.1109/JSTARS.2020.3030304

distribution of representation by minimizing the difference between their covariance matrices. Furthermore, Sun *et al.* extended CORAL to incorporate it directly into deep networks (D-Coral) by constructing a differentiable loss function that minimizes the difference between source and target correlations [23]. In addition, adversarial-based adaptation methods have also achieved outstanding performance. Bousmalis *et al.* employed the original generative adversarial network (GAN) framework to generate fake target data from source data to acquire the relations between domains [24]. Ganin *et al.* proposed a domain adversarial neural network (DANN) for feature alignment by maximizing the domain classification loss [25]. Multiadversarial domain adaptation (MADA) method achieved conditional distribution adaptation by using multiple domain discriminators [26]. Recently, deep learning methods have attracted the most attention in DA field due to its powerful feature representation ability. It can achieve adaptive classification for both domains by learning a shared common feature space.

Deep learning-based DA methods have been successfully applied to the classification of HSIs. Wang *et al.* improved the generalization ability of the classifier on target domain by minimizing the MMD between domains [27]. Li *et al.* minimized the distance between source domain and target domain based on MMD and learned a more discriminative feature space in a two-stage deep DA model [28]. Deng *et al.* employed the adversarial approach and constrained the target embeddings to form similar clusters with the source ones [29]. To preserve the geometric information of original tensors, Qin *et al.* employed a manifold regularization term for core tensors into the optimization process [30]. Liu *et al.* achieved classwise distribution adaptation by using multiple domain classifiers and MMD strategy [31]. A self-attention generative adversarial adaptation network used a GAN equipped with self-attention mechanism to generate high-quality hyperspectral samples, and employed MMD strategy to constrain the generated samples to be more similar to the original ones [32].

For classification of hyperspectral remote sensing images, most DA strategies aim to reduce the distribution shift between domains. Since good classification performance also relies on the discriminability of the features in the common feature space, the DA method that considers both feature alignment and feature discriminability is more desirable. In this article, we focus on associative learning (AL)-based DA [33], which is able to yield features that are both domain-invariant and discriminative. Association means the normalized similarity relationships between two embeddings. When AL is applied to domain adaptation, it is able to learn the cross-domain relations and achieve cross-domain feature alignment. The cross-domain similarity matrix summarizes the relationships between source and target data, and thus an accurate similarity matrix indicates the achievement of class-conditional distribution adaptation between domains. To our best knowledge, the AL-based DA has not been applied for classification of hyperspectral remote sensing images.

AL is achieved by employing the cycle consistency criterion in a two-step transition procedure [33]. The first-step transition indicates the departure walk from source data to target data with the first transition probability matrix. The second-step transition denotes the return walk from target data back to source data with the second transition probability matrix. The two-step transitions produce a cycle that starts from source data and ends up getting back to source data. Since the source data are labeled, we can use the labeled information of the starting and ending source data to determine if a circle is consistent. If a starting source sample belongs to the same class as the ending source sample, the cycle is consistent. Otherwise, the cycle is inconsistent. Since the inconsistent cycles may be produced by the spectral drift or the overlapping spectra between source classes, with a walking loss that is defined to penalize the inconsistent cycles, domain-invariant, and discriminative features can be achieved.

Cross-domain similarity matrix plays an important role in the AL-based DA methods. In this article, we utilize two similarity metrics for the two-step transitions in the cycle. One is calculated by the features as in [34], and the other is obtained by the probability prediction results. The source labels and the target prediction results contain category information and are suitable to describe the relationships between source and target data. With the cycle consistency constraint, the predictions of target data will be aligned with the source labels and tend to have a peak distribution. The AL with feature and prediction-based similarity metrics is referred to as augmented AL (AAL), and the AAL-based DA is denoted as AALDA in this article. It is expected that combining the two similarity metrics can result in a superior DA performance.

The main contributions of this article include the following.

1) We combine the feature-based and probability-prediction-based similarity metrics to describe the relationships between domains, which promotes the feature alignment from different views and obtains a superior DA performance.

2) Applying the proposed cross-domain similarity metrics in the cycle consistency criterion, the AALDA network is able to obtain features that are both domain-invariant and discriminative.

3) To the best of our knowledge, this article is the first attempt to introduce the AL with cycle consistency criterion to DA for hyperspectral remote sensing image classification.

The organization of rest of this article is as follows. Section II presents the proposed AALDA network in detail. Experimental results are discussed in Section III and the conclusion is drawn in Section IV.

II. Proposed DA Method

An unsupervised DA approach attempts to learn a model, which transfers knowledge from a fully labeled source domain to an unlabeled target domain. The labeled source dataset is denoted as $\mathbf{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$, where $\mathbf{x}_i^s \in \mathbf{x}_s$, $\mathbf{y}_i^s$ is the one-hot encoding of the label information and $\mathbf{y}_i^s \in \mathbf{y}_s$ and n_s is the total number of samples on the source dataset. Target dataset is denoted as $\mathbf{D}_t = \{(\mathbf{x}_\mathrm{j}^t)\}_{j=1}^{n_t}$, where $\mathbf{x}_\mathrm{j}^t \in \mathbf{x}_t$ and n_t is the total number of samples on the target dataset. Both the source and target data have the same dimensionality *D* and there are *C* classes in both domains denoted by $\mathbf{\Omega} = [\Omega_1, \ldots, \Omega_C]$.

The flowchart of the proposed approach is shown in Fig. 1. The network contains a feature extractor G_f and a classifier G_c,

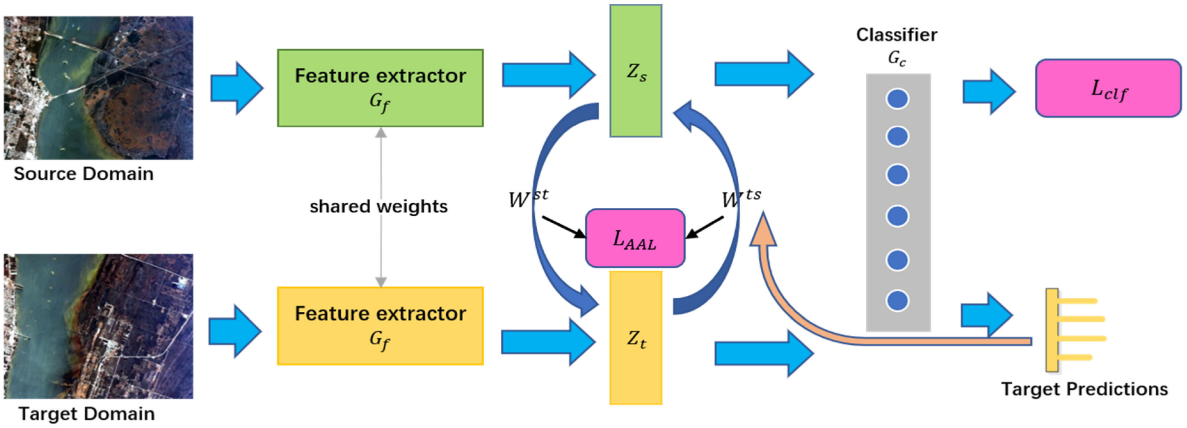


Fig. 1. Flowchart of the augmented AL-based domain adaptation (AALDA) method.

where G_f aims to generate domain-invariant features, and G_c conducts classification and outputs the probability prediction results of the two domains.

The loss function of AALDA is defined as

$$L(\mathbf{X}_s, \mathbf{Y}_s, \mathbf{X}_t; \theta_f, \theta_c) = L_{clf}(\mathbf{X}_s, \mathbf{Y}_s; \theta_f, \theta_c) + \beta_1 L_{walk}(\mathbf{X}_s, \mathbf{X}_t; \theta_f, \theta_c) + \beta_2 L_{visit}(\mathbf{X}_s, \mathbf{X}_t; \theta_f) \quad (1)$$

where L_{clf} represents the classification loss of the labeled source data, L_{walk} denotes the walking loss for the inconsistent cycle, and L_{visit} represents a regularization term to enforce every sample in target domain to be "visited" by source samples. The tradeoff hyper-parameters β_1 and β_2 denote the importance of the walking loss and visiting loss, respectively. The notation θ_f represents the parameter of the feature extractor G_f and θ_c denotes the parameter of the classifier G_c. The feature extractor G_f is trained by all the three losses and G_c is trained by the classification loss and walking loss. It is worth noting that the AALDA network is only composed of full connected layers for pixel-level classification.

A. Source Classification Loss

In DA methods, labeled source data are used to train the model so that it can capture the relevant and more discriminative features. Meanwhile, the feature extractor G_f will generate domain invariant features by utilizing the adaptation loss. Thus, the classifier G_c can be directly used to classify the target features.

The cross-entropy loss is often used as source classification loss in multicategory classification task. Given a labeled source sample $(\mathbf{x}, \mathbf{y})$, its cross-entropy is defined as

$$L_c(\mathbf{p}, \mathbf{y}) = -\sum_{c=1}^{C} y^c \log p^c \quad (2)$$

$$p^c = Softmax(G_c(G_f(\mathbf{x}))_c) = \frac{\exp(G_c(G_f(\mathbf{x}))_c)}{\sum_{i=1}^{C} \exp(G_c(G_f(\mathbf{x}))_i)} \quad (3)$$

where $\mathbf{y}$ is the one-hot label of sample $\mathbf{x}$, $\mathbf{p}$ is the predicted probability output calculated by the output of the classifier G_c, G_c $(G_f(\mathrm{x}))_c$ is the cth output of the classifier on the feature $G_f(\mathrm{x})$, and p^c denotes the probability of $\mathbf{x}$ belonging to the cth class. Thus, the source classification loss is defined as

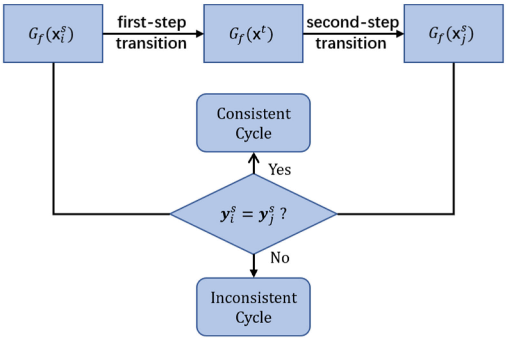


Fig. 2. Illustration of the walking procedure.

$$L_{clf}(\mathbf{X}_s, \mathbf{Y}_s; \theta_f, \theta_c) = \frac{1}{n_s} \sum_{(\mathbf{x}_i,\mathbf{y}_i)\ (\mathbf{X}_s,\mathbf{Y}_s)} L_c(Softmax(G_c(G_f(\mathbf{x}_i))), \mathbf{y}_i) \quad (4)$$

where n_s is the number of samples on the source domain.

B. Walking Loss

We first describe the cycle consistency criterion and the walking loss, and then present the two cross-domain similarity metrics.

1) Cycle Consistency Criterion and Walking Loss: Cycle consistency constraint is imposed to obtain the domain-invariant and discriminative features. Let the network-generated features be denoted as G_f $(\mathbf{X}_s)$ and G_f $(\mathbf{X}_t)$ for source and target data, respectively. A walker starts from G_f $(\mathbf{X}_s)$ to G_f $(\mathbf{X}_t)$ with the first-step transition and then goes back to the source data G_f $(\mathbf{X}_s)$ with the second-step transition. If the starting point and ending point in source domain belong to the same class, the walk is correct and the cycle is consistent. Otherwise, the walk is incorrect or the cycle is inconsistent. The walking procedure is illustrated in Fig. 2.

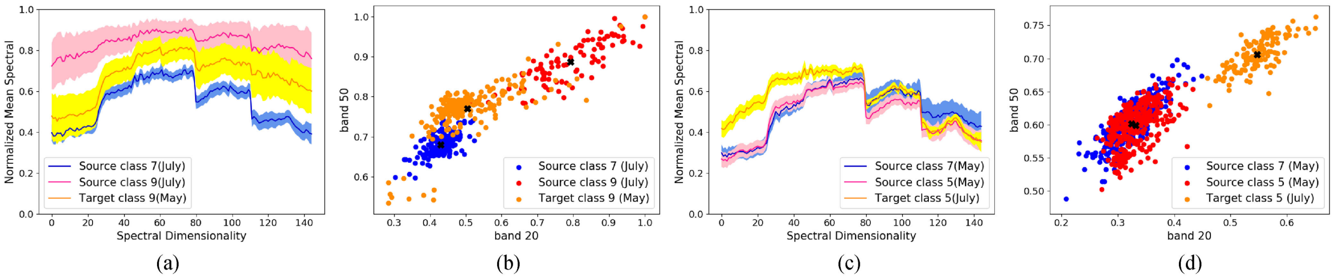


Fig. 3. Illustration of the inconsistent cycles yielded by spectral drift between domains and the overlapping spectra between source classes. (a) Spectral drift of class 9 in BOT July-May data pairs. (b) Scatterplot to show the spectral drift of class 9 in BOT July-May data pair. (c) Overlapping spectra of class 5 and class 7 in source domain. (d) Scatterplot to show the overlapping spectra of class 5 and class 7 in source domain.

Let the matrix $\mathbf{P}^{sts}$ denote the association between source data, with P_{ij}^{sts} denoting the probability of starting at the source sample $G_f\,(\mathbf{x}_i{}^s)$ and ending up at the sample $G_f\,(\mathbf{x}_j{}^s)$. With the two-step transitions, P_{ij}^{sts} is defined as

$$\mathbf{P}_{ij}^{sts} = \sum_{k=1}^{n_t} \mathbf{P}_{ik}^{st}\mathbf{P}_{kj}^{ts} \tag{5}$$

where the superscript "*sts*" in P_{ij}^{sts} denotes the round-trip from source to target and back to source, $\mathbf{P}^{st}$ and $\mathbf{P}^{ts}$ represent two transition probability matrices, $P_{\mathrm{ik}}^{\mathrm{st}}$ represents the transition probability from source sample $G_f\,(\mathbf{x}_i{}^s)$ to target sample $G_f\,(\mathbf{x}_k{}^t)$, and $P_{\mathrm{kj}}^{\mathrm{ts}}$ represents the transition probability from the target sample $G_f\,(\mathbf{x}_k{}^t)$ to source sample $G_f\,(\mathbf{x}_j{}^s)$. It can be seen that the similarity P_{ij}^{sts} between $G_f\,(\mathbf{x}_i{}^s)$ and $G_f\,(\mathbf{x}_j{}^s)$ is expressed by summing their relations to all the target data points. If the two source points are from the same class, there should exist walking path and the P_{ij}^{sts} should be large. Otherwise, they should not be connected via target data and P_{ij}^{sts} should equal to zero. An indicator matrix $\mathbf{T}$ is defined as

$$\mathbf{T}_{ij} = \begin{cases} \frac{1}{n_s^c} & \mathbf{x}_i^s, \mathbf{x}_j^s \in \Omega_c \\ \mathbf{0} & \text{else} \end{cases} \tag{6}$$

where $n_s{}^c$ denotes the number of points in class Ω_c. $\mathbf{T}_{\underline{ij}}$ is a constant variable, which equals to $1/n_s{}^c$ when the two source data points $\mathbf{x}^s{}_i$ and $\mathbf{x}^s{}_j$ are from the same class Ω_c and is equal to zero if they belong to different class. Therefore, the matrix $\mathbf{T}$ is an indicator matrix that accurately characterizes the category relationships between source data points. Then, the matrix $\mathbf{T}$ is used to evaluate the accuracy of matrix $\mathbf{P}^{sts}$. The walking loss of all the source data pairs are then defined as

$$\begin{aligned} L_{walk}(\mathbf{X}^s, \mathbf{X}^t; \theta_f, \theta_c) &= L_c(\mathbf{P}^{sts}, \mathbf{T}) \\ &= \sum_{i,j} -\mathbf{P}_{ij}^{sts} \log \mathbf{T}_{ij} \end{aligned} \tag{7}$$

where L_c is the standard cross-entropy loss function. If the starting and ending source samples are from different classes ($\mathbf{T}_{ij} = 0$) but there is a walk between them ($P_{ij}^{sts}{>}0$), then the cross-entropy loss will be very large.

There are two reasons for yielding inconsistent cycles. One is due to the spectral drift between domains and the other one is yielded by the undistinguishable source classes. By using the walking loss to penalize the inconsistent cycles, an aligned and discriminative feature representation can be achieved.

Algorithm 1: AALDA Approach.

```
Input: Source domain $\mathbf{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$, Target
  domain $\mathbf{D}_t = \{(\mathbf{x}_j^t)\}_{j=1}^{n_t}$, $N$ denotes the batch size, $M$ is
  the max number of iteration, $\mathbf{B}_s^m$ and $\mathbf{B}_t^m$ denote the batch
  data for two domains, $m = 1, \ldots, M$.
Output: The predictions of target data
1. Initialization:
Randomly initializing the parameters of feature extractor
  $G_f$ and the classifier $G_c$.
2. Loading batch data:
    $\mathbf{B}_s^m = \mathrm{BatchSample}(\mathbf{D}_s,\ N)$
    $\mathbf{B}_t^m = \mathrm{BatchSample}(\mathbf{D}_t,\ N)$
3. Training: for $m = 1 \to M$ do
(1) Generating embeddings of both domains:
      $G_f(\mathbf{B}_s^m), G_f(\mathbf{B}_t^m)$
(2) Obtaining the first-step transition matrix:
      $\mathbf{W}^{st} \leftarrow G_f(\mathbf{B}_s^m)^t G_f(\mathbf{B}_t^m), \mathbf{P}^{st} \leftarrow \mathrm{softmax}\,(\mathbf{W}^{st})$
(2) Obtaining the second-step transition matrix:
      $\mathbf{Y}^t \leftarrow \mathrm{softmax}(G_c(G_f(\mathbf{B}_t^m)))$
      $\mathbf{W}_{ji}^{ts} \leftarrow \sum_{k=1}^{C} \mathbf{Y}_j^{tk} \delta\,(\mathbf{y}_i^s, \Omega_k), \mathbf{P}^{ts} \leftarrow \mathrm{softmax}(\mathbf{W}^{ts})$
(4) Obtaining the associations of source
      data:$\mathbf{P}^{sts} \leftarrow \mathbf{P}^{st}\mathbf{P}^{ts}$
(5) Minimizing the objective loss function:
      $\min_{\boldsymbol{\theta}_f, \boldsymbol{\theta}_c} L_{\mathrm{clf}} + \beta_1 L_{\mathrm{walk}} + \beta_2 L_{\mathrm{visit}}$
end for
4. Inference: Predicting target data by the feature extractor
    $G_f$ and the classifier $G_c$.
```

If there exists spectral drift between domains, one target class may become spectrally similar to two different source classes. As a result, two different source classes are falsely associated with the target class, resulting in inconsistent cycles and walking loss. By minimizing the loss using the network, feature alignment can then be achieved. Using the multitemporal Hyperion HSIs on Botswana (BOT), Fig. 3(a)–(b) shows an example to illustrate the spectral drift of class 9 in BOT July–May data pairs. Fig. 3(a) plots the mean spectra with variances of source class 7 and class

9, and plots the target mean spectra with variances from class 9. It can be seen that the target spectra from class 9 are similar to two different source classes due to the spectra drift. We also show the scatterplot of all the points in these classes using two spectral bands in Fig. 3(b). The relationships between all the points can be better visualized. In this case, source data from class 7 and class 9 will be falsely associated with the target data of class 9, and thus inconsistent cycles between the two source classes are produced.

If two source classes have overlapping spectra and are difficult to distinguish, their associations with target data will be similar. If a target class is spectrally similar to one of the source class, it will be similar to the other one as well. Therefore, the two source classes will be falsely associated with the target class. This false association is yielded due to the undistinguishable source classes, and therefore by minimizing the walking loss using the network, the features of the two source classes become discriminative. Fig. 3(c) and (d) illustrates two "similar" source classes and the resulted false association using BOT May–July data pairs. Fig. 3(c) includes the mean spectra of source class 5 and class 7, and target class 5. It can be observed that the two source classes have very similar spectral properties (blue curve and pink curve). Fig. 3(d) shows the scatterplot of these classes with two spectral bands. The overlapping between the two source classes can be clearly observed, and the target samples are close to both of them. Therefore, the two source classes will be falsely associated with the target class.

2) Cross-Domain Similarity Metrics: The two-step transition probability matrices $\mathbf{P}^{st}$ and $\mathbf{P}^{ts}$ play the most important role in the AALDA method, which are calculated from the cross-domain similarity matrices $\mathbf{W}^{st}$ and $\mathbf{W}^{ts}$, respectively. Using the network-generated features, the $\mathbf{W}^{st}$ is expressed as

$$\mathbf{W}_{ij}^{st} = G_f(\mathbf{x}_i^s)^{\mathrm{T}}(G_f(\mathbf{x}_j^t)). \tag{8}$$

Then, the first-step transition $\mathbf{P}^{st}$ can be defined by softmaxing $\mathbf{W}^{st}$ over rows

$$\mathbf{P}_{ij}^{st} = \left(\mathrm{softmax}\left(\mathbf{W}^{st}\right)\right)_{ij} = \frac{\exp\left(\mathbf{W}_{ij}^{st}\right)}{\sum_{j'}\exp\left(\mathbf{W}_{ij'}^{st}\right)}. \tag{9}$$

The second similarity matrix $\mathbf{W}^{ts}$ is obtained from the source labels and the target probability prediction results. Let $\mathbf{Y}_j^t \in \mathrm{R}^{1\times C}$ denotes the probabilistic prediction result of target data $\mathbf{x}_j^t$, and $\mathbf{Y}_j^{tc}$ represents the probability of $\mathbf{x}_j^t$ belonging to the cth class. The similarity between $\mathbf{x}_i^s$ and $\mathbf{x}_j^t$ is expressed as

$$\mathbf{W}_{ji}^{ts} = \sum_{k=1}^{C}\mathbf{Y}_j^{tk}\delta\left(\mathbf{y}_i^s, \Omega_k\right). \tag{10}$$

If $\mathbf{x}_i^s$ belongs to Ω_k, the similarity equals to the probability of $\mathbf{x}_j^t$ belonging to the cth class. The value of $\mathrm{W}_{\mathrm{ji}}^{\mathrm{ts}}$ also equals the inner product of the labeled encoding $\mathbf{y}_i^s$ and probability prediction $\mathbf{y}_j^t$.

Then, the second-step transition probability is

$$\mathbf{P}_{ji}^{ts} = \left(\mathrm{softmax}\left(\mathbf{W}^{ts}\right)\right)_{ji} = \frac{\exp\left(\mathbf{W}_{ji}^{ts}\right)}{\sum_{i'}\exp\left(\mathbf{W}_{ji'}^{ts}\right)}. \tag{11}$$

With these two different similarity matrices, the cycle consistency criterion is augmented to be more robust. And the model will generate domain-invariant and discriminative features by alleviating the spectral drift between domains and the spectral overlapping between source classes.

C. Visiting Loss

The walking loss promotes the correct associations between source data and target data, but it cannot guarantee that every target sample have connections with source data. If some target samples are not associated with source data, these samples will not participate the network training, and the feature extraction is only determined by a portion of target data. Therefore, it is expected that all the target samples are aligned to the corresponding source data. The visiting loss is thus introduced to enforce every target sample to be "visited" by source samples [33], which is defined as

$$L_{visit} = L_c(\mathbf{p}^{visit}, \mathbf{v}) \tag{12}$$

where L_c denotes the cross-entropy loss function. The jth elements in $\mathbf{v} \in \mathrm{R}^{nt\times 1}$ and $\mathbf{p}^{visit} \in \mathrm{R}^{nt\times 1}$ are defined as

$$\mathbf{v}_j = \frac{1}{n_t} \tag{13}$$

$$\mathrm{p}_j^{visit} = \sum_{\mathbf{x}_i\in\mathbf{X}^s}\mathbf{P}_{ij}^{st} \tag{14}$$

where $\mathbf{P}_{ij}^{\mathrm{st}}$ denotes the transition probability from the ith source sample $\mathbf{x}_i^s$ to the jth target sample $\mathbf{x}_j^t$. The p_j^{visit} is the summation of the transition probabilities between $\mathbf{x}_j^t$ to all the source samples. It represents the summarized association with sample $\mathbf{x}_j^t$ to the source domain. The vector $\mathbf{P}^{visit} = [\mathrm{P}_1^{visit}, \mathrm{P}_2^{visit}, \ldots, \mathrm{P}_{n_t}^{visit}] \in \mathrm{R}^{n_t\times 1}$ denotes the associations with all the target samples to source domain. By using the visiting loss function, all the target samples are constrained to be associated with source samples. The procedure of AALDA is summarized in Algorithm 1.

III. Experimental Results and Analysis

Experiments are conducted on two hyperspectral remote sensing images, i.e., Botswana (BOT) dataset and Kennedy Space Center (KSC) dataset, to validate the effectiveness of the proposed approach. The experiments were implemented with the deep learning framework TensorFlow and were executed on NVIDIA GeForce RTX 2080.

A. Experimental Datasets

The BOT dataset was acquired by the NASA EO-1 Hyperion instrument. The Hyperion sensor on EO-1 acquires data at 30 m pixel resolution over a 7.7 km strip in 242 bands covering the 400–2500 nm portion of the spectrum in 10-nm spectral resolution. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features. Three multitemporal images were collected by the Hyperion over the Okavango Delta, Botswana, in May, June, and July 2001. Two of them,

TABLE I
CLASS NAME AND NUMBER OF SAMPLES OF BOT AND KSC IMAGES

BOT					KSC			
ID	Class Name	May	June	July	ID	Class Name	KSC1	KSC2
1	Water	297	361	185	1	Scrub	761	422
2	Primary Floodplain	437	308	96	2	Willow Swamp	243	180
3	Riparian	448	303	164	3	Cabbage Palm Hammock	256	431
4	Firescar	354	335	186	4	Cabbage Palm/Oak	252	132
5	Island Interior	337	370	131	5	Slash Pine	161	166
6	Woodlands	357	324	169	6	Oak/Broadleaf Hammock	229	274
7	Savanna	330	342	171	7	Hardwood Swamp	105	248
8	Short Mopane	239	299	152	8	Graminoid Marsh	431	453
9	Exposed Soils	215	229	96	9	Salt Marsh	419	156
					10	Water	927	1392
Total		3014	2871	1350			3784	3854

TABLE II
OA (OA%) OF DIFFERENT UNSUPERVISED DA ALGORITHMS

Data sets		SO	DANN	MADA	DAN	D-Coral	ST	ADA	AALDA
BOT	May-June	87.86	88.23	90.54	89.91	90.32	92.84	92.61	**93.79**
	June-May	79.21	78.15	80.19	84.97	87.36	85.60	86.89	**91.14**
	May-July	84.56	86.79	89.97	87.48	89.48	90.86	90.95	**91.25**
	July-May	66.77	68.69	72.93	73.78	75.30	82.45	82.16	**83.59**
	June-July	94.81	94.47	94.93	91.84	94.49	95.27	95.13	**95.30**
	July-June	89.75	89.40	89.80	90.13	91.47	94.83	93.66	**94.50**
KSC	KSC1-KSC2	62.78	65.12	71.79	64.82	68.56	67.02	71.89	**73.53**
	KSC2-KSC1	61.26	61.00	74.85	63.26	61.99	69.39	71.54	**75.03**

June and July images, have the same field-of-view. And the other one, the May image, was acquired over a neighboring but overlapping area. The three images with 1476 × 256 pixels include nine identified classes which have been shown in Table I . We can pick out two of the three images as source and target data, respectively. Thus, six data pairs are available for our DA experiments. The pseudo-color image and label information in May, June, and July are shown in Fig. 4.

The NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) instrument acquired data over the Kennedy Space Center (KSC), Florida, on March 23, 1996. AVIRIS acquires data in 224 bands of 10 nm width with center wavelengths from 400–2500 nm. The KSC data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. After removing water absorption and low SNR bands, 176 bands are used for the analysis. KSC data contain two spatially disjoint images, where one image includes a protected wildlife area and the other one has experienced anthropogenic impacts. The two images are named as Area1 and Area2 of KSC and are denoted as KSC1 and KSC2. Both of the two images contain ten classes which have been listed in Table I. Discrimination of land cover for this environment is difficult due to the similarity of spectral signatures for certain vegetation types. The pseudocolor image and label information are shown in Fig. 5.

For BOT dataset, we chose multitemporal images for source and target data. Thus, there are six data pairs that were used for DA experiments, which are named as “May–June,” “June–May,” “May–July,” “July–May,” “June–July,” and “July–June.” In the scheme of “May–June,” there are 3014 samples and 2871 samples in source data and target data, respectively.

For KSC dataset, images at different locations are chosen for domain adaptation. We name two data pairs as “KSC1–KSC2”

TABLE III
KAPPA COEFFICIENTS OF DIFFERENT UNSUPERVISED DA ALGORITHMS

Data sets		SO	DANN	MADA	DAN	D-Coral	ST	ADA	AALDA
BOT	May-June	0.865	0.869	0.893	0.886	0.891	0.920	0.922	**0.923**
	June-May	0.762	0.748	0.776	0.830	0.857	0.838	0.855	**0.901**
	May-July	0.825	0.850	0.886	0.858	0.881	0.897	0.884	**0.897**
	July-May	0.628	0.647	0.695	0.704	0.721	0.800	0.772	**0.808**
	June-July	0.942	0.937	0.943	0.908	0.938	0.947	0.936	**0.949**
	July-June	0.880	0.881	0.885	0.889	0.904	0.944	0.926	**0.938**
KSC	KSC1-KSC2	0.545	0.577	0.673	0.575	0.616	0.601	0.670	**0.675**
	KSC2-KSC1	0.556	0.545	0.698	0.564	0.555	0.648	0.646	**0.705**

TABLE IV
CLASSIFICATION ACCURACY OF EACH CLASS OF BOT "MAY-JUNE" DATA

ID	Class	SO	DANN	MADA	DAN	D-Coral	ST	ADA	AALDA
1	Water	100.00	100.00	100.00	100.00	100.00	100.00	100.00	**100.00**
2	Primary Floodplain	83.44	94.48	96.85	98.38	99.35	96.75	98.38	**98.70**
3	Riparian	92.07	89.11	90.83	90.43	87.13	92.08	**93.40**	91.42
4	Firescar	100.00	99.70	99.43	100.00	99.10	100.00	100.00	**100.00**
5	Island Interior	96.49	99.73	98.92	98.11	100.00	98.65	**99.73**	99.19
6	Woodlands	43.83	62.04	59.07	50.93	62.65	60.80	55.86	**61.73**
7	Savanna	99.12	92.11	97.95	93.57	82.75	99.12	99.12	**99.42**
8	Short Mopane	95.65	86.62	98.63	98.33	94.31	99.33	**99.67**	99.33
9	Exposed Soils	68.12	59.83	63.14	70.31	79.91	86.90	82.10	**86.03**
	OA	87.86	88.23	90.54	89.91	90.32	92.84	92.61	**93.79**
	AA	89.19	87.07	89.43	88.84	89.47	92.53	92.06	**93.20**
	kappa	0.865	0.869	0.893	0.886	0.891	0.920	0.922	**0.923**

TABLE V
RUNNING TIME(S) OF ONE STEP TRAINING OF DIFFERENT METHODS ON TWO DATASETS

Dataset	SO	DANN	MADA	DAN	D-Coral	ST	ADA	AALDA
May-June	0.012	0.029	0.049	0.026	0.024	0.047	0.055	0.071
KSC1-KSC2	0.016	0.030	0.051	0.027	0.025	0.048	0.058	0.074

and "KSC2–KSC1." For "KSC1–KSC2" data pairs, we use the 3784 samples and 3854 samples as source data and target data, respectively.

For HSIs, spectral drift often occurred in the spatially/temporally separate images, due to different conditions such as illumination, topography, soil moisture, and vegetation composition. Using BOT May-June, May-July, and KSC1-KSC2 data pairs, Fig. 6 illustrates the spectral drift between domains and its influence on classification. Fig. 6(a) shows the mean spectra of source classes 2 and 5 and target class 2. The spectral drift can be observed between the pink curve (source class 2) and the orange curve (target class 2). Moreover, the target spectra become closer to another source class (class 5, blue curve), and thus the target data from class 2 may be falsely predicted as class 5 if the classifier is only trained by source labeled data. Fig. 6(b) and (c) obtain the similar observations that spectral drift of a category may yield misclassification and the DA can be applied to solve the problem.

B. Implementation Detail

The compared DA networks include the multikernel MMD strategy-based DAN [21], CORAL-based network [23], class-wise centroid alignment-based semantic transfer (ST) [35], adversarial learning-based DANN [25], multiple domain discriminators-based MADA network [26], and the associative domain adaptation (ADA) network [33].

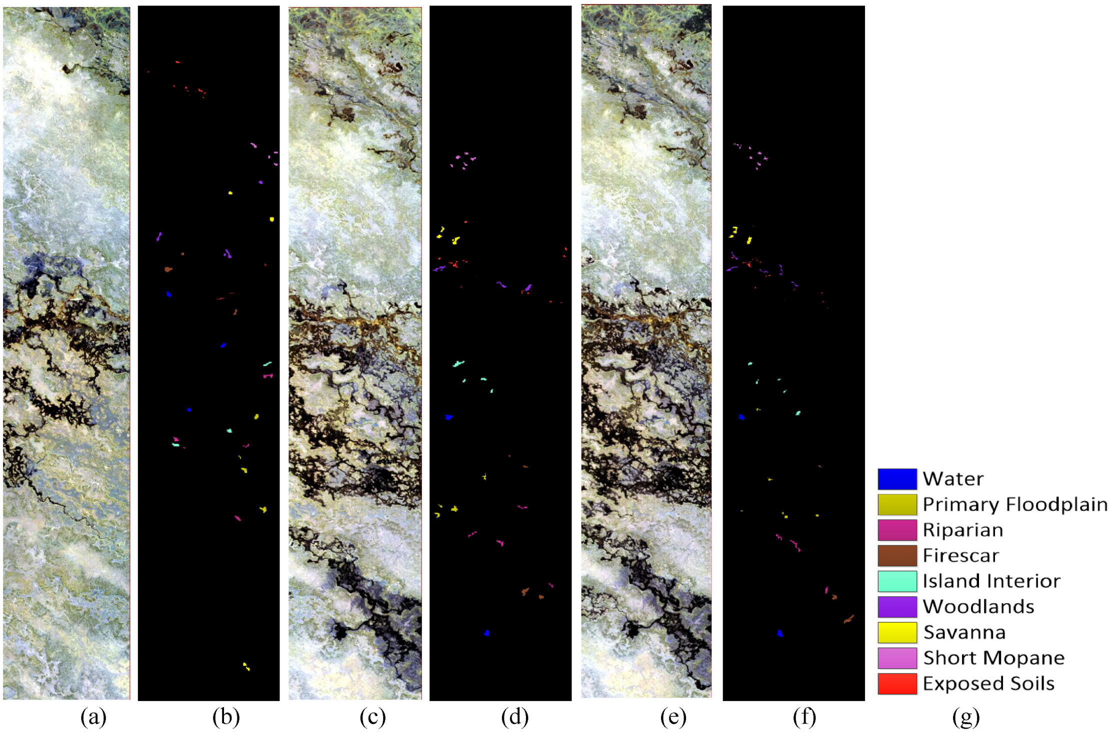


Fig. 4. Hyperion images of BOT in May (a), June (c), and July (e). (b) Ground truth of image in May. (d) Ground truth of image in June. (f) Ground truth of image in July. (g) Class legend.

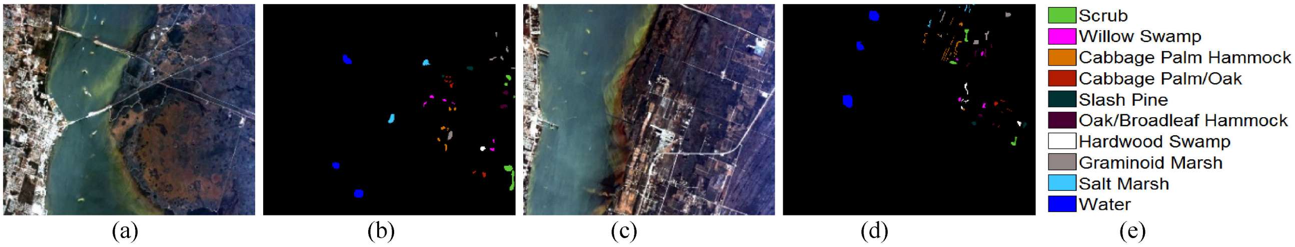


Fig. 5. RGB images of (a) KSC1 and (c) KSC2. Labeled data of (b) KSC1 and (d) KSC2. (e) Class legend.

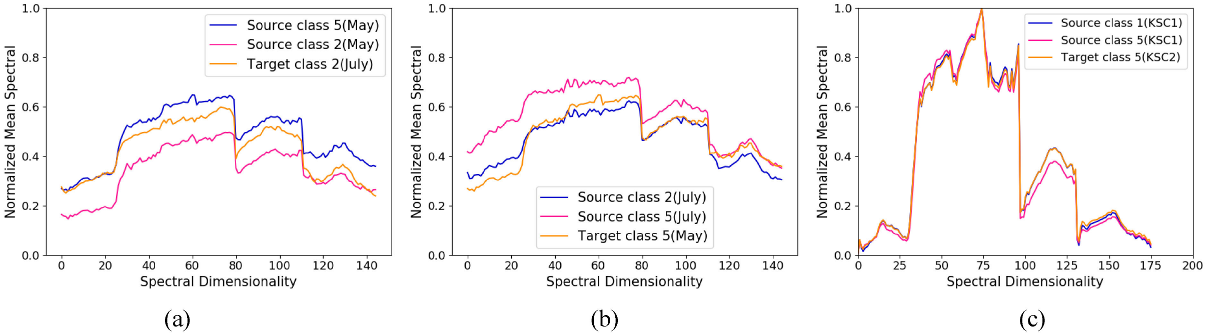
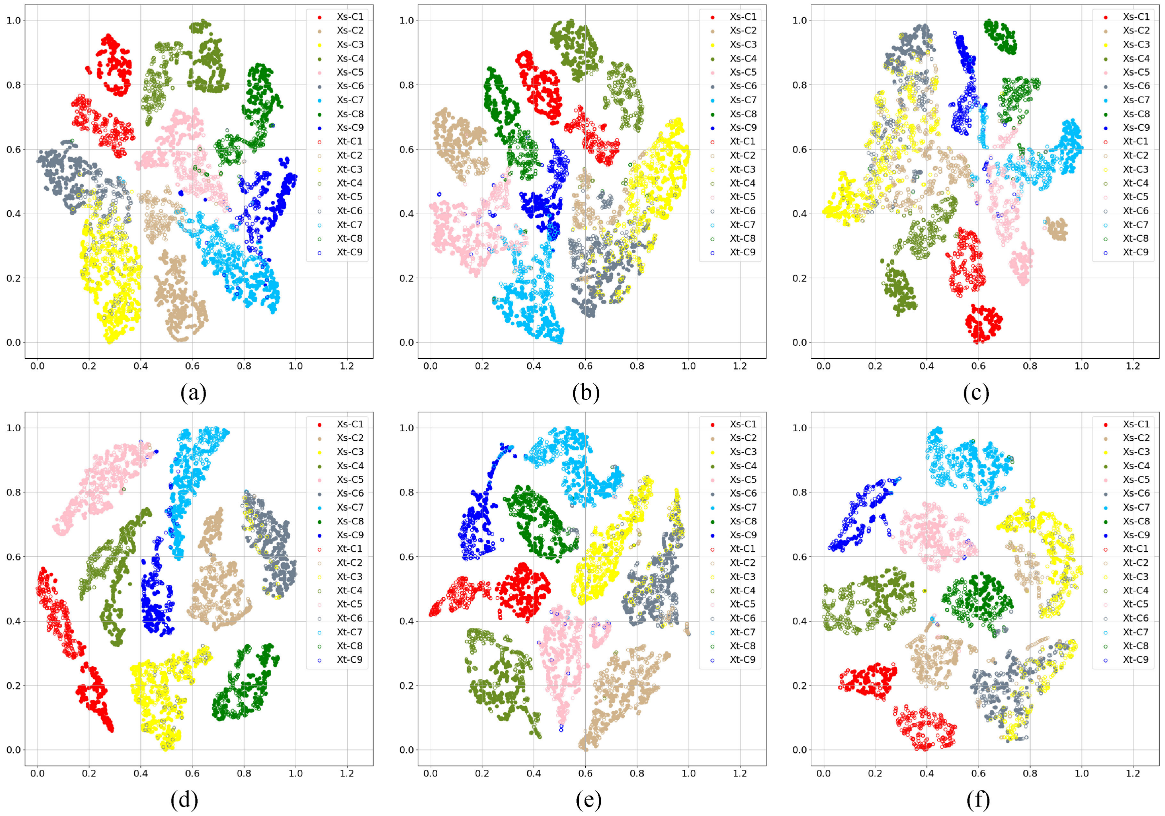


Fig. 6. Illustration of how the spectral drift effects the classification performance. (a) Spectral drift of "May-June." (b) Spectral drift of "July-May." (c) Spectral drift of "KSC1-KSC2."



(a) (b) (c)

(d) (e) (f)

Fig. 7. Alignment performance of AALDA for BOT “May-June,” “June–May,” and “July-May” data. (a) “May-June” data pair before alignment. (b) “June–May” data pair before alignment. (c) “July-May” data pair before alignment. (d) “May-June” data pair after alignment. (e) “June–May” data pair after alignment. (f) “July-May” data pair after alignment.

In the proposed AALDA, the feature extractor is composed of three fully connected layers (128, 64, and 32 units in sequence). The classifier is composed of a softmax output layer with dimensionality being set to the number of classes. Each hidden layer employs the Leaky ReLU [36] activation function. Note that we did not use BN [37] or dropout [38] to avoid affecting the performance of adaptation. All the compared methods used the feature extractor and classifier with the same architecture.

Data preprocessing is applied before training to normalize each spectral band with standard normal distribution *N(0,1)*. Thus, the overall distribution shift across domains will be partly reduced. The same data preprocessing method was also applied to the compared approaches.

We used Adam [39] to optimize the network in all experiments. In AALDA, the cycle associations may be affected if a mini-batch does not include all classes. Thus, we propose to select a larger batch so that all classes can be included. The mini-batch size is set as 128 in AALDA and all the comparison approaches. The learning rate is set 0.001 initially and decay by a factor of 0.33 every 4000 steps. The model trained on all datasets can converge in less than 800 epochs. The tradeoff hyper-parameters β_1 and β_2 control the ratio between walking loss and the visiting loss. We tested the ratio in the range 1:2, 1:1, 2:1, and 3:1. Meanwhile, we introduced a factor α to alter the value of β_1 and β_2. The factor α was selected from 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, and 4.0. In the compared approaches, DAN and deep CORAL have a weight for the MMD loss and CORAL loss, respectively, DANN and MADA has a weight for the domain confusion loss, ST contains a parameter for the semantic transfer loss, and ADA has a ratio in the associate loss. Their recommended values were used.

C. Alignment Performance of AALDA

The alignment performance of AALDA is shown in Fig. 7. We used t-SNE [40] method to visualize the high-dimensional embedding in 2-D. As shown in Fig. 7, source embeddings and target embeddings are represented by dots and pentagrams, respectively, and different classes shown in different colors. In Fig. 7(a)–(c), we plotted the embedding generated by the feature extractor which is trained on labeled source domain only. As a contrast, we plotted the embedding obtained from the feature extractor, which employed the AALDA training scheme in Fig. 7(d), (e), and (f). There are biases between each source class and target class. By employing the AALDA method, the distribution of each source class and each corresponding target class are well aligned. The comparison verified the outstanding alignment performance of the proposed AALDA method.

D. Experiment Results and Comparison

The evaluation criteria used in this article include overall accuracy (OA), average accuracy (AA), and kappa coefficients

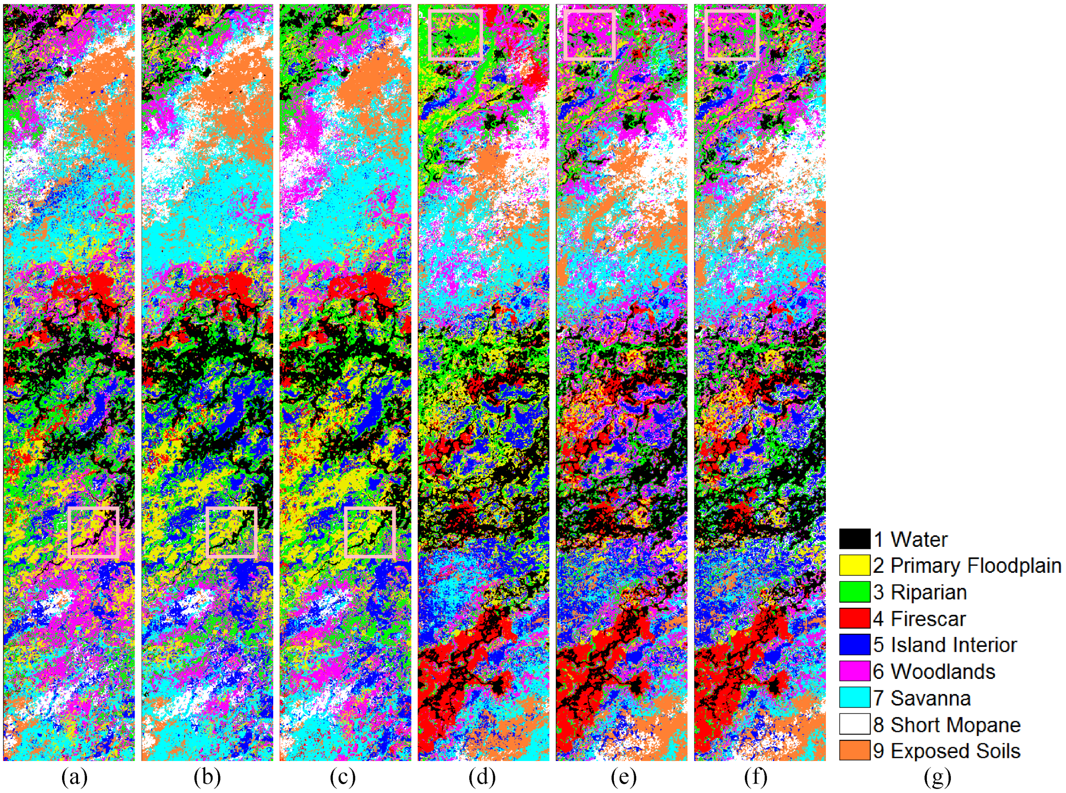


Fig. 8. Classification results of the target image in BOT "June-May" and "June-July" data pairs. (a) SO result without any adaptation for "June-May" data pair. (b) AALDA result for "June-May" data pair. (c) Reference obtained by using target labeled data as training data for "June-May" data pair. (d) SO result without any adaptation for "June-July" data pair. (e) AALDA result for "June-July" data pair. (f) Reference obtained by using target labeled data as training data for "June-July" data pair. (g) Class legend.

(Kappa). OA is a ratio between the number of correctly classified testing samples and the number of testing samples. AA is the AA of all the classes. Kappa is a statistic that is used to measure the agreement of classification for all classes. The OA and Kappa of all above algorithms on BOT and KSC dataset can be found in Tables II and III, respectively. The unsupervised adaptation results on eight data pairs as adaptability may vary across different transfer tasks. In addition, "Source only" (SO) refers to method which trained only on source domain. The gap on OA between SO and other algorithms represent the transferability of these algorithms. The results verify that our method achieves promising performance and outperforms the compared methods. It can be seen that the SO scheme without any transfer strategy still achieved satisfactory performance on some data pairs with small spectral drift, such as BOT "June-July" and "July-June." The SO scheme obtained low accuracies on some data pairs with big spectral drift, such as BOT "July-May," "KSC1-KSC2," and "KSC2-KSC1." Almost all the compared methods can obtain higher accuracies than the SO scheme, which demonstrates their positive transfer learning ability. The proposed AALDA outperforms DANN, MADA, DAN, and D-Coral. It demonstrates that enforcing the cycle consistency between domains obtains better feature alignment than other methods which only align the marginal distribution of two domains. AALDA obtained almost 5%–15% improvement on the accuracy with respect to the SO scheme. AALDA utilizes both the feature-based similarity and the similarity based on probabilistic prediction results to construct the roundtrip transitions. The AALDA performs better than ADA, demonstrating the effectiveness of introducing the prediction-based similarity measurement.

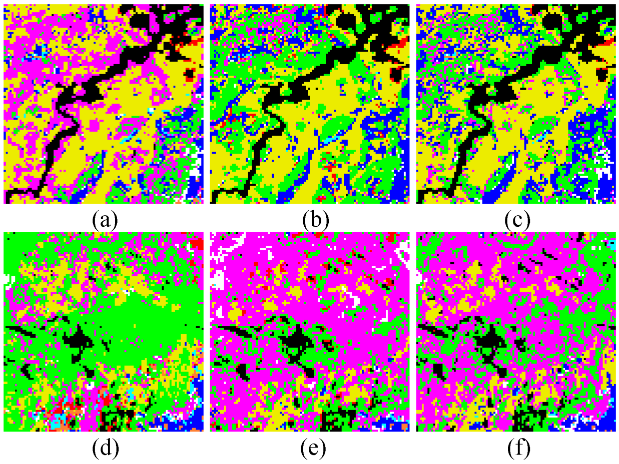
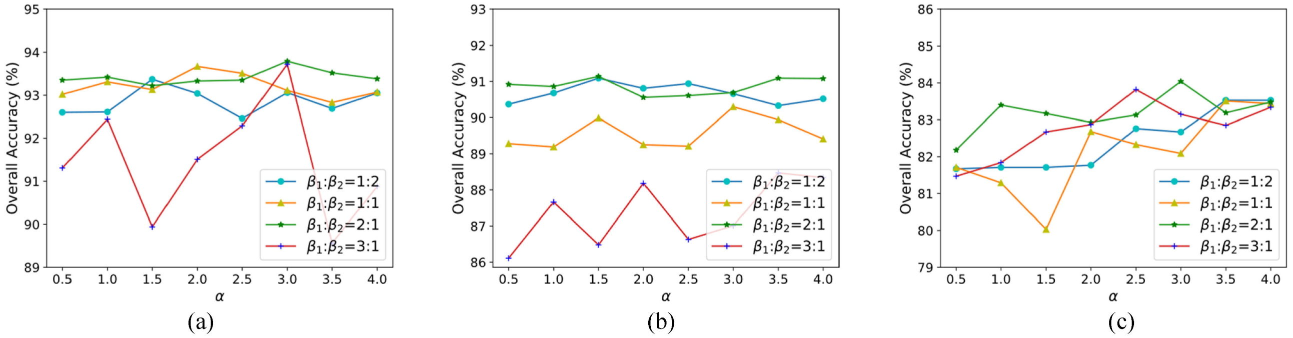


Fig. 9. Classification results of local regions. (a) SO result without any adaptation for "June-May" wetland area. (b) AALDA result for "June-May" wetland area. (c) Reference obtained by using target labeled data as training data for "June-May" wetland area. (d) SO result without any adaptation for "June-July" upland area. (e) AALDA result for "June-July" upland area. (f) Reference obtained by using target labeled data as training data for "June-July" upland area.



(a) (b) (c)

Fig. 10. Sensitivity analysis of parameters in our adaptation using BOT data pairs. (a) BOT "May-June." (b) BOT "June-May." (c) BOT "July-May."

The performance of classification on per-class basis shows transfer ability in a detailed way. We picked BOT "May-June" data pair to conduct this experiment and the results showed in Table IV. It can be seen that the AALDA algorithm outperformed on almost all classes, suggesting that the AALDA can transfer knowledge in different classes. Moreover, the AALDA achieved the most improvement on some classes which are different to be classified correctly, such as class 6 and class 9.

E. *Classification Results of the Whole Image by the AALDA*

We selected BOT "June-May" and "June-July" data pairs to illustrate the performance of the classification on the whole images. The classification results of the whole images are shown in Fig. 8. The results of the SO scheme without adaptation strategy for BOT "June-May" and "June-July" are shown in Fig. 8(a) and (d), respectively. Fig. 8(b) and (e) shows the classification results of the proposed AALDA approach. Due to the lack of ground truth for the whole image, we train a classifier with target labeled data and regard the output of the classifier as "reference." As shown in Fig. 8, we can observe that the results of AALDA are very similar to the "reference," which demonstrate the transfer ability of the AALDA approach.

To observe the details, we chose two local regions from the two data pairs in Fig. 8 and enlarged in Fig. 9. The BOT dataset includes two major eco-systems: upland and wetland [41]. We selected a wetland and an upland in the BOT "June-May" and "June-July" data pairs, separately. For the BOT "June-May" data pair, the most part of the selected wetland local region are samples from class 1 (Water, black), class 2 (Primary Floodplain, yellow), and class 3 (Riparian, light green). As shown in Fig. 9(a)–(c), the most of samples in class 3 (Riparian, light green) were misclassified as class 6 (Woodlands, purple) in the results of SO scheme without any adaptation. Fig. 9(b) shows the results of AALDA and is similar to the "reference" in Fig. 9(c). Similarly, Fig. 9(d)–(f) are results for an upland local area, which mainly contains samples from class 6 (Woodlands, purple). But many samples were misclassified as class 3 (Riparian, light green). After AALDA, most of samples in class 3 (Riparian, light green) can be correctly classified. It demonstrates that the AALDA provided satisfactory classification performance for the HSIs.

F. *Sensitivity Analysis of the Parameters*

In the proposed AALDA, two hyper-parameters β_1 and β_2 control the importance of the walking loss and the visiting loss. We set the ratio $\beta_1 : \beta_2$ as 1:2, 1:1, 2:1, and 3:1. Meanwhile, α alters the value of β_1 and β_2, and can balance the weight between adaptation loss and the source classification loss. The performance of different hyper-parameters on the BOT "May-June," "June-May," and "July-May" data pairs is shown in Fig. 10. Specifically, when the ratio $\beta_1 : \beta_2$ is set as 2:1, AALDA can achieve superior performance. Therefore, we suggest setting the ratio $\beta_1 : \beta_2$ as 2:1 in the AALDA approach.

G. *Computation Cost*

The running time (training) of each method on BOT May-June data pair and KSC1-KSC2 data pair are shown in Table V. We implemented all the experiments with the TensorFlow deep learning library on a workstation equipped with an Intel Core i7-8700 CPU (16GB RAM) and a Nvidia GTX 2080 GPU with 8 GB memory. As shown in Table V, MADA and ST that employ conditional distribution adaptation is more time-consuming than DANN, DAN, and D-Coral that perform margin distribution adaptation. ADA and AALDA cost more than the other methods due to the calculation of the transition probability matrix. AALDA costs the most since two different transition probability matrixes need to be computed, but the computational time is still acceptable.

IV. Conclusion

In this article, we proposed a novel method for cross-domain classification of remote sensing images by utilizing two similarity metrics to describe the relations between domains. The proposed AALDA approach employs the criterion of cycle consistency to generate features that are both domain-invariant and discriminative. Experiments on hyperspectral remote sensing images demonstrate its effectiveness compared to other unsupervised DA methods. Moreover, the proposed AALDA approach outperforms ADA, indicating that the proposed prediction-based similarity metric is able to promote consistent cycles. In the future work, we may use other novel similarity metric to conduct the cycle consistency. In addition, attention mechanism can also be introduced to DA network [32].

Acknowledgment

The authors would like to thank Professor M. Crawford at Purdue University for providing the BOT and KSC data used in this study.

References

[1] J. Jiang, J. Ma, Z. Wang, C. Chen, and X. Liu, "Hyperspectral image classification in the presence of noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 851–865, Feb. 2019.

[2] 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.

[3] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul.2019.

[4] J. Peng, W. Sun, and Q. Du, "Self-paced joint sparse representation for the classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1183–1194, Feb. 2019.

[5] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, Jun. 2019.

[6] T. Tian, Z. Pan, X. Tan, and Z. Chu, "Arbitrary-oriented inshore ship detection based on multi-scale feature fusion and contextual pooling on rotation region proposals," *Remote Sens.*, vol. 12, Jan. 2020, Art. no. 339.

[7] 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.

[8] 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.

[9] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.

[10] S. Ben-David, J. Blitzer, K. Crammer, and A. Kulesza, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1/2, pp. 151–175, 2010.

[11] Y. Zhang, T. Liu, M. Long, and M. I. Jordan, "Bridging theory and algorithm for domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7404–7413.

[12] S.J Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[13] 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.

[14] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 4918–4927.

[15] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, "Semisupervised manifold alignment of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7708–7720, Dec. 2014.

[16] S. Wang, L. Zhang, W. Zuo, and B. Zhang, "Class-specific reconstruction transfer learning for visual recognition across domains," *IEEE Trans. Image Process.*, vol. 29, pp. 2424–2438, 2020.

[17] I. Jhuo, D. Liu, D. T. Lee, and S. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2168–2175.

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, A. Smola, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, Mar. 2012.

[19] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach." in *Proc. IEEE Int. Conf. Comput. Vision*, Barcelona, Spain, 2011, pp. 999–1006.

[20] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv*:1412.3474.

[21] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, France, 2019, pp. 97–105.

[22] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," in *Proc. Domain Adaptation Comput. Vision Appl.*, 2017, pp. 153–171,.

[23] B. Sun and S. Kate, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 443–450.

[24] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 95–104.

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, and F. Laviolette. "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.

[26] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018. [Online]. Available: https://arxiv.org/abs/1809.02176

[27] 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.

[28] Z. Li, X. Tang, W. Li, C. Wang, C. Liu, and J. He, "A two-stage deep domain adaptation method for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 7, pp. 1054–1073, Mar. 2020.

[29] B. Deng, J. Sen, and D. Shi, "Deep metric learning-based feature embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1422–1435, Feb. 2020.

[30] 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.

[31] Z. Liu, L. Ma, and Q. Du, "Class-wise distribution adaptation for unsupervised classification of hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, to be published.

[32] W. Zhao, X. Chen, J. Chen, and Y. Qu, "Sample generation with self-attention generative adversarial adaptation network (SaGAAN) for hyperspectral image classification," *Remote Sens.*, vol. 12, pp. 843, Jan. 2020.

[33] P. Haeusser, A. Mordvintsev, and D. Cremers, "Associative domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, 2017, pp. 2784–2792.

[34] O. Sener, H O. Song, A. Saxena, and S. Savarese, "Learning transferrable representations for unsupervised domain adaptation," in *Proc. 30th Int. Conf. Neural Inf. Proc. Syst.*, Spain, 2016, pp. 2118–2126.

[35] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, 2018, pp. 5423–5432.

[36] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, vol. 30, no. 1, Jun. 2013, pp. 3.

[37] S. Ioffe and S. Christian, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[38] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[39] D P. Kingma and J Ba, "Adam: A method for stochastic optimization," 2014, *arXiv*:1412.6980.

[40] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, Beijing, China, 2014, pp. 647–655.

[41] A. L. Neuenschwander, "Remote Sensing of Vegetation Dynamics in Response to Flooding and Fire in the Okavango Delta, Botswana," Ph.D. dissertation, Dept. Aerosp. Eng., Univ. Austin, TX, USA pp. 2007.

Min Chen (Student Member, IEEE) received the B.S. degree in information engineering from Wuhan University of Technology, Wuhan, China, in 2018. He is currently working toward the M.S degree in electronics and communication engineering with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, China.

His research interests include pattern recognition, computer vision and hyperspectral data analysis.

Li Ma (Member, IEEE) received the B.S. degree in biomedical engineering and M.S. degree in pattern recognition and intelligent from Shandong University, Jinan, China, in 2004 and 2006, respectively, and the Ph.D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology, Wuhan, China, in 2011.

During 2008–2010, she was a visiting scholar with Purdue University, Indiana, USA. She also visited Mississippi State University, Mississippi, USA, for five months, in 2018. She is currently an Associate Professor with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan. Her research interests include hyperspectral data analysis, pattern recognition, and remote sensing applications.

Wenjin Wang (Student Member, IEEE) received the B.S. degree in communication engineering from the China University of Geosciences, Wuhan, China, in 2018, where she is currently working toward the M.S. degree in information and communication engineering with School of Mechanical Engineering and Electronic Information.

Her research interests include pattern recognition, computer vision, and hyperspectral data analysis.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland at Baltimore, Baltimore, MD, USA, in 2000.

She is the Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA, and also an Adjunct Professor with the College of Surveying and Geo-informatics, Tongji University, Shanghai, China. Her research interests include hyperspectral remote sensing image analysis and applications, pattern classification, data compression, and neural networks.

Dr. Du is a Fellow of the SPIC International Society for Optics and Photonics. She was a recipient of the 2010 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society. She was the Chair of the Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition from 2010 to 2014. She served as the Co-Chair for the Data Fusion Technical Committee of the IEEE GRSS from 2009 to 2013. She was the General Chair for the fourth IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing held at Shanghai, in 2012. She has served as an Associate Editor for the IEEE JOURNAL OF TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2011–2015), the *Journal of Applied Remote Sensing* (2014–2015), and the IEEE SIGNAL PROCESSING LETTERS (2012–2015). She also organized several international workshops and journal special issues on remote sensing image processing and analysis.