




# Hyperspectral Image Classification Based on Domain Adaptation Broad Learning

Haoyu Wang, Xuesong Wang , *Member, IEEE*, C. L. Philip Chen , *Fellow, IEEE*,  
and Yuhu Cheng , *Member, IEEE*

**Abstract**—Hyperspectral images (HSI) are widely applied in numerous fields for their rich spatial and spectral information. However, in these applications, we always face the situation that the available labeled samples are limited or absent. Therefore, we propose an HSI classification method based on domain adaptation broad learning (DABL). First, according to the importance of the marginal and conditional distributions, the maximum mean discrepancy is used in mapped features to adapt these distributions between source and target domains. Meanwhile the manifold regularization is added to maintain the manifold structure of the input HSI data. Second, to further reduce the distribution difference and maintain manifold structure, the domain adaptation and manifold regularization are added to the output layer of DABL. Finally, the output weights can be easily calculated by the ridge regression theory. Experimental results on three real HSI datasets demonstrate the effectiveness of our proposed DABL.

**Index Terms**—Broad learning, classification, domain adaptation, hyperspectral image (HSI).

## I. INTRODUCTION

**H**YPERSPECTRAL images (HSIs) contain rich spectral features and spatial information about surface objects on the earth [1], which is widely applied to the fields of environmental monitoring, crop monitoring, mineral exploration, etc. [2]. These successful applications often greatly rely on appropriate data processing approaches, such as target detection, physical or chemical parameter retrieval, and classification [3]. HSI classification is the common task for data processing. Many methods have been proposed to improve the classification accuracy of HSI [4]–[7], such as extreme learning machine [1], support vector machine (SVM) [8], [9], and neural network [10]. These supervised classification methods often require a large number of

labeled samples to obtain the satisfactory classification accuracy [11]. However, it is expensive and difficult to collect the labeled data on HSIs [12]. To address this concern, some machine learning methods have been proposed. For example, active learning is a technique to train a classifier with a small quantity of labeled samples, enabling the classifier to actively select representative unlabeled samples [13]. Semisupervised learning is an effective approach to utilize a large amount of unlabeled data with some labeled samples for image classification [14]. Different from directly reducing labeling costs and extracting information from unlabeled samples, domain adaptation is a particular form of transfer learning [15], utilizing the samples from related domain (source domain) to solve problems for another domain (target domain) [16],[17]. When there are insufficient samples in the target domain, the same or similar labeled samples from the source domain can be used. For example, Long *et al.* [18] proposed a joint distribution adaptation (JDA) that can jointly adapt both the marginal and conditional distributions. Chen *et al.* [19] proposed to reduce the domain distribution difference between the source and target domains using extreme learning machine framework, named domain space transfer ELM (DST-ELM). Ganin and Lempitsky [20] proposed a domain-adversarial neural network (DANN) to select transferable features from different domains by introducing the adversarial mechanism into deep transfer network.

The domain adaptation technique was successfully applied to HSI classification. Sun *et al.* [21] proposed to simultaneously minimize the maximum mean discrepancy (MMD) [22],[23] and the structural risk item of SVMs. Xia *et al.* [24] divided the feature space of the source and target domains into several disjoint feature subspaces, and then exploited transfer component analysis (TCA) to obtain integrated features of each subspace. In [25], Sun *et al.* designed the transfer sparse subspace analysis to learn some sparse subspaces across domains, thus the features from both domains in the subspaces were aligned. Li *et al.* [26] proposed to learn the best projection matrices for heterogeneous domains in a sparse subspace, and then utilized the canonical correlation analysislike regularization to design an appropriate classifier. In recent years, deep domain adaptation has been successfully applied to HSI classification and acquired high classification accuracy. Riz *et al.* [27] trained a classifier with the domain invariant features acquired by stacked denoising autoencoders. Zhou and Prasad [28] extracted the discriminative features for two domains with deep convolutional recurrent neural networks, and then the features were aligned with each other

Manuscript received January 27, 2020; revised April 22, 2020 and May 23, 2020; accepted June 4, 2020. Date of publication June 9, 2020; date of current version June 18, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61976215, Grant 61772532, Grant 61751202, and Grant U1813203. (Corresponding author: Yuhu Cheng.)

Haoyu Wang, Xuesong Wang, and Yuhu Cheng are with the Engineering Research Center of Intelligent Control for Underground Space Ministry of Education, Xuzhou Key Laboratory of Artificial Intelligence and Big Data, and the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China (e-mail: wanghaoyucumt@163.com; wangxuesongcumt@163.com; chengyuhu@163.com).

C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau 999078, China (e-mail: philip.chen@ieee.org).

Digital Object Identifier 10.1109/JSTARS.2020.3001198

layer-by-layer in the common subspaces. However, these deep network-based domain adaptation algorithms generally require complicated structure and a time-consuming training process [29].

Recently, Chen and Liu [30] first proposed a novel broad learning system (BLS) based on random vector functional-link neural network (RVFLNN) consisting of merely three parts: mapped feature (MF), enhancement node (EN), and output layer. The BLS has the following advantages.

- 1) BLS has a simple structure with only three parts.
- 2) The network weights of BLS are calculated with the ridge regression. Deep learning methods utilize gradient descending, which requires more times of iterations. Therefore, the training speed of BLS is efficient.
- 3) The MF mapped to EN achieves broad expansion and feature enhancement, which makes BLS has a strong function approximation capability.

Feng and Chen [31] combined the Takagi–Sugeno fuzzy system with BLS, which achieved a satisfactory accuracy in classification. Jin *et al.* [32] merged the manifold learning into BLS to classify images. For the past few years, BLS has also been well applied to HSI. Kong *et al.* [33] merged the class-probability structure into BLS to obtain a semisupervised learning version, which achieves a high accuracy in HSI classification. Kong *et al.* [34] fine-tuned the weights of MF and EN with the graph-regularized sparse autoencoder, which maintained the manifold structure of original data. However, both the aforementioned two HSI classification methods cannot help to improve the HSI classification accuracy by utilizing vast quantities of labeled samples in related domains. Moreover, if the classifier trained with labeled samples from source domain is used to classify the samples from target domain directly, the classification accuracy will be low due to the distribution difference between the source and target domains. Therefore, we propose an HSI classification method based on domain adaptation broad learning (DABL). In summary, the main contributions of our work are as follows.

- 1) We propose a DABL method by introducing transfer learning technology to the traditional BLS and apply it to HSI classification. The DABL can realize the unsupervised classification of target domain HSI by only using the labeled HSI data from the source domain.
- 2) The distribution difference between the source and target domains is adapted with MMD based on a distribution importance parameter, where the importance of marginal and conditional distributions is evaluated according to the A-distance (A-d) between two domains.
- 3) Not only in the MF layer, but also in the output layer of DABL, both the MMD and manifold regularization are utilized. Thus, the distribution difference between the source and target domains can be further reduced and the manifold structure of HSI data can be well maintained.

The rest of this article is organized as follows. In Section II, related work, including BLS and MMD, is briefly introduced. Details of the proposed DABL for HSI classification are presented in Section III. Experiments results are reported in Section IV, followed by a conclusion in Section V.

## II. RELATED WORK

### A. Broad Learning System

The BLS is a new type of flat network, which is designed based on the idea of RVFLNN [30]. The structure of the BLS is showed in Fig. 1, which can be viewed as a three-layer feedforward neural network. The workflow of BLS can be illustrated as follows. First, the original inputs  $\mathbf{X}$  are mapped to feature nodes via random weights. Suppose there are  $m$  groups of feature nodes, the  $i$ th group MF is [30]

$$\mathbf{Z}_i = \phi(\mathbf{X}\mathbf{W}_{ei} + \beta_{ei}), \quad i = 1, \dots, m \quad (1)$$

where  $\mathbf{W}_{ei}$  and  $\beta_{ei}$  are the connecting weight and bias from input to MF,  $\phi(\cdot)$  is the activation functions of MF. Then, the MF is randomly mapped to EN for broad expansion and the  $j$ th group EN is [30]

$$\mathbf{H}_j = \sigma(\mathbf{Z}^n \mathbf{W}_{hj} + \beta_{hj}), \quad j = 1, \dots, e \quad (2)$$

where  $\mathbf{W}_{hj}$  and  $\beta_{hj}$  are connecting weight and bias from MF to EN,  $\sigma(\cdot)$  is the activation function. Finally, both MF and EN are connected to the output layer, and the network output is [30]

$$\mathbf{O} = [\mathbf{Z} | \mathbf{H}] \mathbf{W}^O \quad (3)$$

where  $\mathbf{W}^O = [\mathbf{Z} | \mathbf{H}]^+ \mathbf{O}$ . The objective function of BLS is [30]

$$\min_{\mathbf{W}^O} \|\mathbf{O} - \mathbf{Y}\|_2^2 + \delta \|\mathbf{W}^O\|_2^2 \quad (4)$$

where  $\mathbf{Y}$  is the label of input  $\mathbf{X}$  and  $\delta$  is the regularization parameter.

### B. Maximum Mean Discrepancy

MMD is an effective nonparametric distance metric [22]. In the field of domain adaptation, MMD is generally used to reduce the distribution difference between domains and learn the domain invariant features. Suppose there are two probability distributions  $s$  and  $t$ ,  $\mathcal{H}$  is the high-dimensional reproducing kernel Hilbert space, and  $\vartheta(\cdot)$  is a nonlinear mapping function in  $\mathcal{H}$ , then MMD is defined as

$$D_f(s, t) = \sup_{\|\vartheta\|_{\mathcal{H}} \leq 1} \|E_{\mathbf{X}_s \sim s}[\vartheta(\mathbf{X}_s)] - E_{\mathbf{X}_t \sim t}[\vartheta(\mathbf{X}_t)]\|_{\mathcal{H}}^2 \quad (5)$$

where  $E_{\mathbf{X}_s \sim s}[\cdot]$  is the mathematical expectation about distribution  $s$ ,  $\mathcal{H}$  is a set of functions defined with  $\|\vartheta\| \leq 1$  as the unit sphere. If and only if  $s = t$ , we have  $D_f(s, t) = 0$ . Given observations  $D_s = \{\mathbf{X}_{s(i)}\}_{i=1}^M$  and  $D_t = \{\mathbf{X}_{t(j)}\}_{j=1}^N$  drawn independently and identically distributed from  $s$  and  $t$ , respectively, the empirical estimate of MMD is

$$D_f(D_s, D_t) = \left\| \frac{1}{M} \sum_{i=1}^M \vartheta(\mathbf{X}_{s(i)}) - \frac{1}{N} \sum_{j=1}^N \vartheta(\mathbf{X}_{t(j)}) \right\|_{\mathcal{H}}^2. \quad (6)$$

## III. HSI CLASSIFICATION BASED ON DABL

The flowchart of the proposed DABL for HSI classification is shown in Fig. 2, which mainly contains five steps, which are as follows.

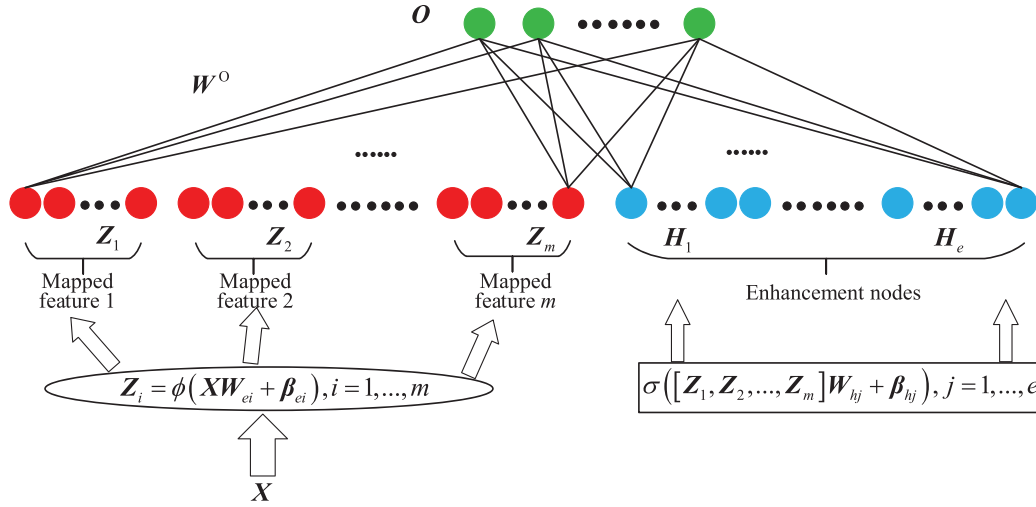


Fig. 1. Structure of BLS.

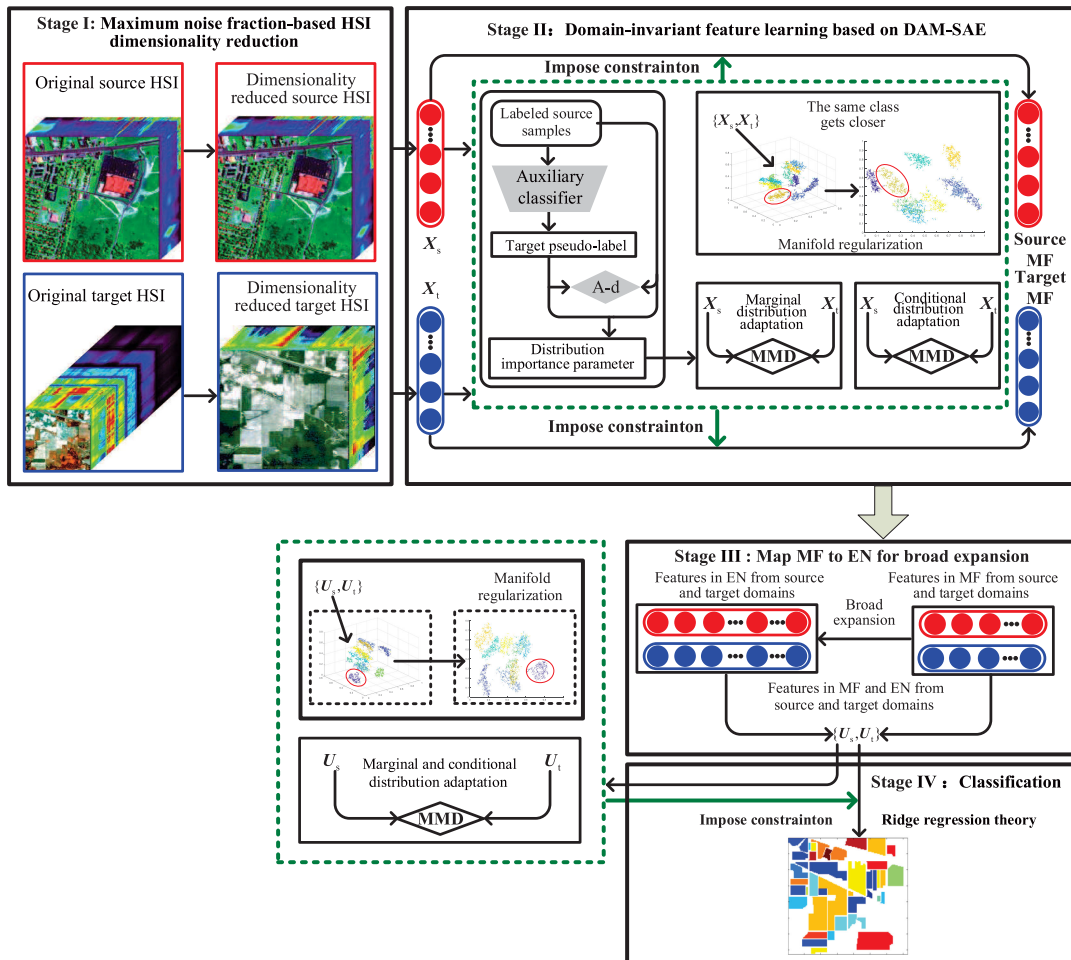


Fig. 2. Flowchart of DABL for HSI classification.

- 1) The maximum noise fraction (MNF) is applied to the original HSI to remove noise and reduce dimensionality.
- 2) The target domain pseudolabels are obtained according to the auxiliary classifier trained on the source domain, based on which the distribution importance parameter can

- be calculated by using A-d to measure the importance of the marginal and conditional distributions.
- 3) The marginal and conditional distribution adaptation terms are, respectively, constructed by MMD based on the distribution importance parameter, which are together

with the manifold regularization term to constrain the stacked autoencoder (SAE). Thus, the domain-invariant features of the source and target HSIs, i.e., source MFs and target MFs, can be extracted with the domain adaptation manifold SAE (DAM-SAE).

- 4) The source and target MFs are mapped to ENs with randomly generated weights for broad expansion. Furthermore, the features in MF and EN from the source and target domains are connected and fed to the output layer of DABL.
- 5) According to the objective function of DABL with distribution adaptation and manifold regularization, the output layer weights can be calculated with the ridge regression theory.

#### A. MNF-Based HSI Dimensionality Reduction

Different bands of HSI are usually highly correlated, especially for adjacent bands, and there is noise in the original HSI [35]. Therefore, MNF is applied to the original HSI to reduce the dimensionality and eliminate noise [36]. MNF can find a linear transformation matrix  $\mathbf{W}_M$  by maximizing the signal-to-noise ratio of HSI.  $\mathbf{W}_M$  can be calculated by

$$\operatorname{argmax}_{\mathbf{W}_M} \frac{\mathbf{W}_M^T \mathbf{C}_V \mathbf{W}_M}{\mathbf{W}_M^T \mathbf{C}_N \mathbf{W}_M} = \operatorname{argmax}_{\mathbf{W}_M} \frac{\mathbf{W}_M^T \mathbf{C}_S \mathbf{W}_M}{\mathbf{W}_M^T \mathbf{C}_N \mathbf{W}_M} - 1. \quad (7)$$

Assuming  $\mathbf{S} = \mathbf{V} + \mathbf{N}$ , where  $\mathbf{S}$ ,  $\mathbf{V}$ , and  $\mathbf{N}$  are original data, uncorrelated signal, and noise matrix, respectively, we get  $\operatorname{cov}(\mathbf{S}) = \mathbf{C}_S = \mathbf{C}_V + \mathbf{C}_N$ , where  $\mathbf{C}_S$ ,  $\mathbf{C}_V$ , and  $\mathbf{C}_N$  are covariance matrix of  $\mathbf{S}$ ,  $\mathbf{V}$ , and  $\mathbf{N}$ .  $\mathbf{W}_M$  are eigenvectors from  $F$  largest eigenvalues of  $\mathbf{C}_N^{-1} \mathbf{C}_S$ , and  $F$  denotes the number of MNF principal components. The dimension reduced  $\mathbf{X}$  as the input of the model is obtained as

$$\mathbf{X} = \mathbf{W}_M^T \mathbf{S}. \quad (8)$$

Note that  $\mathbf{C}_S$  is obtained by calculating the covariance of the samples and  $\mathbf{C}_N$  can be obtained by the minimum/maximum autocorrelation factors method.

#### B. Domain-Invariant Feature Learning Based on DAM-SAE

In the network of original BLS, Chen and Liu [30] map the input data with the weights fine-tuned by SAE to MF. The connection weights from MF to EN are randomly generated. However, neither random generation nor SAE fine-tuned can reduce the distribution difference between training and testing samples. Many domain adaptation methods map the data of source domain and target domain to a subspace, and then reduce the distribution difference by minimizing the MMD of the source and target domain features in the subspace [37]. Therefore, based on SAE, we reduce the marginal and conditional distribution divergences between two domains by adding domain adaptation regularization terms.

Suppose there are MNF-based HSI samples  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_s+n_t}\} \in \mathbb{R}^{(n_s+n_t) \times d}$ , where  $\mathbf{X}_s = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_s}\} \in \mathbb{R}^{n_s \times d}$  and  $\mathbf{X}_t = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_t}\} \in \mathbb{R}^{n_t \times d}$  are the samples from source and target domains,  $n_s$  and  $n_t$  are numbers of source samples and target samples,  $d$  is the dimension of samples.  $\mathbf{X}$  is

mapped to MF by  $d^M$  groups of weights  $\mathbf{A}_i$ , then we can obtain

$$\mathbf{Z}_i = \mathbf{X} \mathbf{A}_i \quad (9)$$

where  $\mathbf{Z}_i \in \mathbb{R}^{(n_s+n_t) \times G^M}$  is the  $i$ th group MFs and  $G^M$  represents the feature dimension of each group. Similar to SAE, the optimization equation here is

$$\operatorname{argmin}_{\mathbf{A}_i} \|\mathbf{X} \mathbf{A}_i - \mathbf{Z}_i\|_2^2 + \lambda \|\mathbf{A}_i\|_1 \quad (10)$$

where  $\lambda$  denotes the regularization parameter. To reduce the distribution difference between the source and target domains, simultaneously, we adapt both marginal and conditional distributions between MFs of the two domains

$$\begin{aligned} & \operatorname{argmin}_{\mathbf{A}_i} \|\mathbf{X} \mathbf{A}_i - \mathbf{Z}_i\|_2^2 + \lambda \|\mathbf{A}_i\|_1 + \theta_1 D_f(P_s, P_t) \\ & + \theta_2 \sum_{c=1}^C D_f(Q_s, Q_t) \end{aligned} \quad (11)$$

where  $\theta_1$  and  $\theta_2$  represent parameters of marginal and conditional distribution regularization terms, respectively.  $c \in \{1, 2, 3, \dots, C\}$  is the class index, and  $C$  is the number of classes.  $D_f(P_s, P_t)$  is used to align the marginal probability distribution of the source and target domains, and  $\sum_{c=1}^C D_f(Q_s, Q_t)$  is used to align the conditional probability distribution.

When a large difference exists between datasets, the marginal probability distribution adaptation becomes important [38]. In contrast, when the datasets are similar, the conditional probability distribution adaptation becomes important [28], [38]. By borrowing the idea of dynamic distribution alignment [39], we exploit the distribution importance parameter  $\mu$  to measure the importance of two distributions, and the entire domain adaptation regularization term can be expressed as

$$D_f = (1 - \mu) D_f(P_s, P_t) + \mu \sum_{c=1}^C D_f(Q_s, Q_t). \quad (12)$$

$\mu \rightarrow 0$  means the distribution distance between the source and target domains is large. At this time, the marginal distribution adaptation becomes important. When  $\mu \rightarrow 1$  means the distribution distance between the source and target domains is small. It is important to align each class, so the conditional distribution adaptation becomes important. A-d can be used to measure the similarity between two distributions [40]. A linear classifier is built to distinguish the loss between two data, and the A-d can be represented as

$$d_A(D_s, D_t) = 2(1 - 2\varepsilon(h)) \quad (13)$$

where  $\varepsilon(h)$  is the loss of the classifier. For the marginal probability distribution difference, we directly use (13) to calculate the A-d  $d_M$  between  $D_s$  and  $D_t$ . For the conditional distribution difference, we use k-nearest neighbor (KNN) algorithm to train an auxiliary classifier with source samples. After that, the auxiliary classifier is used to obtain the pseudolabel on target domain. Finally, the A-d  $d_c$  for the  $c$ th class can be calculated



as  $d_c = d_A(D_s^{(c)}, D_t^{(c)})$ . Thus,  $\mu$  can be obtained by

$$\mu \approx 1 - \frac{d_M}{d_M + \sum_{c=1}^C d_c}. \quad (14)$$

Equation (11) can be transformed as

$$\begin{aligned} & \underset{\mathbf{A}_i}{\operatorname{argmin}} \|\mathbf{X}\mathbf{A}_i - \mathbf{Z}_i\|_2^2 + \lambda \|\mathbf{A}_i\|_1 \\ & + \alpha \left[ (1 - \mu) D_f(P_s, P_t) + \mu \sum_{c=1}^C D_f(Q_s, Q_t) \right] \end{aligned} \quad (15)$$

where  $\alpha$  is the domain adaptation parameter.

However, mapping the input data to the MF only through SAE ignores the intrinsic structure of the input data, such as manifold structure. Therefore, to maintain the same manifold structure of MF as the input data, a manifold regularization term is added during input mapping to the MF. According to the manifold assumption [41], if two data points  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are close to each other in the original data distribution, the MFs  $\mathbf{z}_i$  and  $\mathbf{z}_j$  should also be close to each other. A manifold regularization term is added to (15), thus we have

$$\begin{aligned} & \underset{\mathbf{A}_i}{\operatorname{argmin}} \|\mathbf{X}\mathbf{A}_i - \mathbf{Z}_i\|_2^2 + \lambda \|\mathbf{A}_i\|_1 \\ & + \alpha \left[ (1 - \mu) D_f(P_s, P_t) + \mu \sum_{c=1}^C D_f(Q_s, Q_t) \right] \\ & + \beta \sum_{i=1}^{n_s+n_t} \sum_{j=1}^{n_s+n_t} a_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \end{aligned} \quad (16)$$

where  $\beta$  is the manifold regularization parameter, and  $a_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\psi^2)$ . The Lagrangian expression of (16) is

$$\begin{aligned} & \underset{\mathbf{A}_i}{\operatorname{argmin}} \|\mathbf{X}\mathbf{A}_i - \mathbf{Z}_i\|_2^2 + \lambda \|\mathbf{A}_i\|_1 + \alpha \operatorname{tr}(\mathbf{A}_i^T \mathbf{X}^T \mathbf{M} \mathbf{X} \mathbf{A}_i) \\ & + \beta \operatorname{tr}(\mathbf{Z}_i^T \mathbf{L} \mathbf{Z}_i) \end{aligned} \quad (17)$$

where the domain adaptation regularization term of (16) can be represented as

$$\begin{aligned} & (1 - \mu) D_f(P_s, P_t) + \mu \sum_{c=1}^C D_f(Q_s, Q_t) \\ & = \operatorname{tr}(\mathbf{A}_i^T \mathbf{X}^T \mathbf{M} \mathbf{X} \mathbf{A}_i) \end{aligned} \quad (18)$$

where the marginal distribution  $D_f(P_s, P_t)$  is

$$\begin{aligned} & D_f(P_s, P_t) \\ & = \left\| \frac{1}{n_s} [1 \ 1 \ \cdots \ 1]_{1 \times n_s} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_s} \end{bmatrix}_{n_s \times 1} \mathbf{A} \right. \\ & \quad \left. - \frac{1}{n_t} [1 \ 1 \ \cdots \ 1]_{1 \times n_t} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_s} \end{bmatrix}_{n_t \times 1} \mathbf{A} \right\|^2 \end{aligned}$$

$$\begin{aligned} & = \operatorname{tr} \left( \frac{1}{n_s^2} \mathbf{1} \mathbf{X}_s \mathbf{A} (\mathbf{1} \mathbf{X}_s \mathbf{A})^T - \frac{1}{n_s n_t} \mathbf{1} \mathbf{X}_s \mathbf{A} (\mathbf{1} \mathbf{X}_t \mathbf{A})^T \right. \\ & \quad \left. - \frac{1}{n_s n_t} \mathbf{1} \mathbf{X}_t \mathbf{A} (\mathbf{1} \mathbf{X}_s \mathbf{A})^T + \frac{1}{n_t^2} \mathbf{1} \mathbf{X}_t \mathbf{A} (\mathbf{1} \mathbf{X}_t \mathbf{A})^T \right) \\ & = \operatorname{tr} \left( \frac{1}{n_s^2} \mathbf{1} \mathbf{X}_s \mathbf{A} \mathbf{A}^T \mathbf{X}_s^T \mathbf{1}^T - \frac{1}{n_s n_t} \mathbf{1} \mathbf{X}_s \mathbf{A} \mathbf{A}^T \mathbf{X}_t^T \mathbf{1}^T \right. \\ & \quad \left. - \frac{1}{n_s n_t} \mathbf{1} \mathbf{X}_t \mathbf{A} \mathbf{A}^T \mathbf{X}_s^T \mathbf{1}^T + \frac{1}{n_t^2} \mathbf{1} \mathbf{X}_t \mathbf{A} \mathbf{A}^T \mathbf{X}_t^T \mathbf{1}^T \right) \\ & = \operatorname{tr} \left[ \mathbf{A}^T \left( \frac{1}{n_s^2} \mathbf{X}_s^T \mathbf{1}^T \mathbf{1} \mathbf{X}_s - \frac{1}{n_s n_t} \mathbf{X}_t^T \mathbf{1}^T \mathbf{1} \mathbf{X}_s \right. \right. \\ & \quad \left. \left. - \frac{1}{n_s n_t} \mathbf{X}_s^T \mathbf{1}^T \mathbf{1} \mathbf{X}_t - \frac{1}{n_t^2} \mathbf{X}_t^T \mathbf{1}^T \mathbf{1} \mathbf{X}_t \right) \mathbf{A} \right] \\ & = \operatorname{tr} \left( \mathbf{A}^T \begin{bmatrix} \mathbf{X}_s^T & \mathbf{X}_t^T \end{bmatrix} \begin{bmatrix} \frac{1}{n_s^2} \mathbf{1}^T \mathbf{1} & -\frac{1}{n_s n_t} \mathbf{1}^T \mathbf{1} \\ -\frac{1}{n_s n_t} \mathbf{1}^T \mathbf{1} & \frac{1}{n_t^2} \mathbf{1}^T \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_t \end{bmatrix} \mathbf{A} \right) \\ & = \operatorname{tr}(\mathbf{A}^T \mathbf{X}^T \mathbf{M}_0 \mathbf{X} \mathbf{A}) \end{aligned} \quad (19)$$

where  $\mathbf{M}_0$  is

$$(\mathbf{M}_0)_{ij} = \begin{cases} \frac{1}{n_s^2}, & \mathbf{x}_i, \mathbf{x}_j \in D_s \\ \frac{1}{n_t^2}, & \mathbf{x}_i, \mathbf{x}_j \in D_t \\ -\frac{1}{n_s n_t}, & \text{otherwise} \end{cases}$$

Similar to the marginal distribution (19), the conditional distribution can be rewritten as

$$\begin{aligned} D_f(Q_s, Q_t) & = \sum_{c=1}^C \left\| E[f(\mathbf{z}_s^{(c)})] - E[f(\mathbf{z}_t^{(c)})] \right\|_{\mathcal{H}}^2 \\ & = \operatorname{tr}(\mathbf{A}^T \mathbf{X}^T \mathbf{M}_1 \mathbf{X} \mathbf{A}) \end{aligned} \quad (20)$$

where  $\mathbf{M}_1 = \sum_{c=1}^C \mathbf{M}_c$ ,  $\mathbf{M}_c$  can be rewritten as

$$(\mathbf{M}_c)_{ij} = \begin{cases} \frac{1}{(n_s)_c^2}, & \mathbf{x}_i, \mathbf{x}_j \in D_s^{(c)} \\ \frac{1}{(n_t)_c^2}, & \mathbf{x}_i, \mathbf{x}_j \in D_t^{(c)} \\ -1, & \begin{cases} \mathbf{x}_i \in D_s^{(c)}, \mathbf{x}_j \in D_t^{(c)} \\ \mathbf{x}_i \in D_t^{(c)}, \mathbf{x}_j \in D_s^{(c)} \end{cases} \\ 0, & \text{otherwise} \end{cases}$$

Let  $\mathbf{M} = (1 - \mu) \mathbf{M}_0 + \mu \sum_{c=1}^C \mathbf{M}_c$ , then we have  $(1 - \mu) D_f(P_s, P_t) + \mu \sum_{c=1}^C D_f(Q_s, Q_t) = \operatorname{tr}(\mathbf{A}_i^T \mathbf{X}^T \mathbf{M} \mathbf{X} \mathbf{A}_i)$  where  $(n_s)_c$  is the number of  $c$ th class source domain samples, and  $(n_t)_c$  is the number of  $c$ th class target domain samples. The manifold regularization term in (16) is

$$\sum_{i=1}^{n_s+n_t} \sum_{j=1}^{n_s+n_t} a_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 = \operatorname{tr}(\mathbf{Z}_i^T \mathbf{L} \mathbf{Z}_i) \quad (21)$$

where  $L$  is the Laplace matrix that can be obtained by constructing a KNN graph.  $L = D - W$ , where  $D$  is a diagonal degree matrix and  $D_{ii} = \sum_{j=1}^{n_s+n_t} W_{ij}$ .  $W$  is a similarity matrix. Equation (21) can be solved by the alternating direction method of multipliers (ADMM) [40]. One-norm is the nonconvex function and an auxiliary variable  $O$  is introduced here. Thus, (16) can be written as

$$\begin{aligned} & \underset{A_i}{\operatorname{argmin}} \|XA_i - Z_i\|_2^2 + \lambda \|O\|_1 \\ & + \alpha \operatorname{tr}(A_i^T X^T M X A_i) + \beta \operatorname{tr}(Z_i^T L Z_i) \\ & \text{s.t. } A_i - O = 0. \end{aligned} \quad (22)$$

The Lagrangian expression of (16) is

$$\begin{aligned} J = & \underset{A_i}{\operatorname{argmin}} \|XA_i - Z_i\|_2^2 + \lambda \|O\|_1 \\ & + \alpha \operatorname{tr}(A_i^T X^T M X A_i) + \beta \operatorname{tr}(Z_i^T L Z_i) \\ & + \rho \omega^T (A_i - O) + \frac{\rho}{2} \|A_i - O\|_2^2 \end{aligned} \quad (23)$$

where  $\rho > 0$  is a constant. In the light of ADMM,  $A_i$ ,  $O$ , and  $\omega$  are updated alternatively, updating one variable at a time and fixing the other two variables.

1)  $A_i$  can be obtained by solving the following formula:

$$A_i^{(k+1)} = \underset{A_i}{\operatorname{argmin}} J(A_i, O, \omega). \quad (24)$$

By calculating the derivative of  $J$  with respect to  $A_i$  and setting it to zero, we can obtain

$$A_i^{(k+1)} = \frac{X^T Z_i + \rho(O^{(k)} - \omega^{(k)})}{X^T X + \rho I + X^T (\alpha M + \beta L) X}. \quad (25)$$

2)  $O$  can be updated by

$$O^{(k+1)} = S_{\lambda/\rho}(A_i^{(k+1)} + \omega^{(k)}) \quad (26)$$

where  $S_k(\cdot)$  is the soft threshold operation, and  $S_k(\cdot)$  can be calculated by

$$S_k(g) = \begin{cases} g - k, & g > k \\ 0, & |g| \leq k \\ g + k, & g < -k \end{cases} \quad (27)$$

where  $k$  is the artificially defined threshold, such as  $10^{-3}$ .

3) The update formula  $\omega$  is

$$\omega^{(k+1)} = \omega^{(k)} + (A_i^{(k+1)} - O^{(k+1)}). \quad (28)$$

The aforementioned three steps are performed alternately until convergence or reaching the predefined number of iterations, and then the required  $A_i$  can be obtained. Then,  $Z_i$  can be calculated by

$$Z_i = X A_i. \quad (29)$$

Features in MF from source domain and target domain can be represented as  $Z_i^s = X_s A_i$  and  $Z_i^t = X_t A_i$ . MFs are mapped

to EN with randomly generated EN weights  $W^E$  to achieve broad expansion by

$$H = \sigma(ZW^E) \quad (30)$$

where  $Z = [Z_1, Z_2, \dots, Z_{d^E}]$ ,  $\sigma(\cdot)$  is tansig function here, and  $H \in \mathbb{R}^{(n_s+n_t) \times d^E}$  are features of EN.  $d^E$  is the number of nodes in EN.

Features in EN from source domain and target domain can be represented by  $H_s = \sigma(Z_s W^E)$  and  $H_t = \sigma(Z_t W^E)$ .  $Z_s$  and  $H_s$  are features in MF and EN from source domain.  $Z_t$  and  $H_t$  are features in MF and EN from target domain.

### C. HSI Classification Based on DABL

To further reduce the distribution difference between source and target domains, and to maintain the manifold structure of HSI, domain adaptation and manifold regularization terms are added into the objective function. The objective function of DABL can be expressed as

$$\begin{aligned} & \underset{W}{\operatorname{argmin}} \|[Z_s|H_s]W - Y_s\|_2^2 + \delta \|W\|_2^2 \\ & + \eta \left[ (1 - \mu) D_f^{\text{ME}}(P_s, P_t) + \mu \sum_{c=1}^C D_f^{\text{ME}}(Q_s, Q_t) \right] \\ & + \gamma \sum_{i=1}^{n_s+n_t} \sum_{j=1}^{n_s+n_t} b_{ij} \|y_i - y_j\|_2^2 \end{aligned} \quad (31)$$

where  $D_f^{\text{ME}}(P_s, P_t)$  denotes the marginal distribution difference between source and target domains in MF and EN,  $\sum_{c=1}^C D_f^{\text{ME}}(Q_s, Q_t)$  represents the conditional distribution difference between the two domains in MF and EN, and  $\sum_{i=1}^{n_s+n_t} \sum_{j=1}^{n_s+n_t} b_{ij} \|y_i - y_j\|_2^2$  is the manifold regularization term.

Let  $U_s = [Z_s|H_s]$ ,  $U_t = [Z_t|H_t]$ , and  $U = [Z|H]$ , the Lagrangian expression of (31) is

$$\begin{aligned} R = & \underset{W}{\operatorname{argmin}} \|U_s W - Y_s\|_2^2 + \delta \|W\|_2^2 \\ & + \eta \operatorname{tr}(W^T U^T M U W) + \gamma (W^T U^T L U W) \end{aligned} \quad (32)$$

where  $\eta$  is the domain adaptation parameter, and  $\delta$  and  $\gamma$  are the regularization parameters.

Similar to the calculation of  $A$ , the output layer weight  $W$  can be obtained by

$$W = \frac{U_s^T Y_s}{U_s^T U_s + \delta I + \eta U^T M U + \gamma U^T L U}. \quad (33)$$

The output of DABL can be obtained as

$$Y = U W. \quad (34)$$

Furthermore, the target domain classification result  $Y_t$  can be calculated as

$$Y_t = U_t W. \quad (35)$$

The steps of HSI classification based on DABL are summarized as follows.

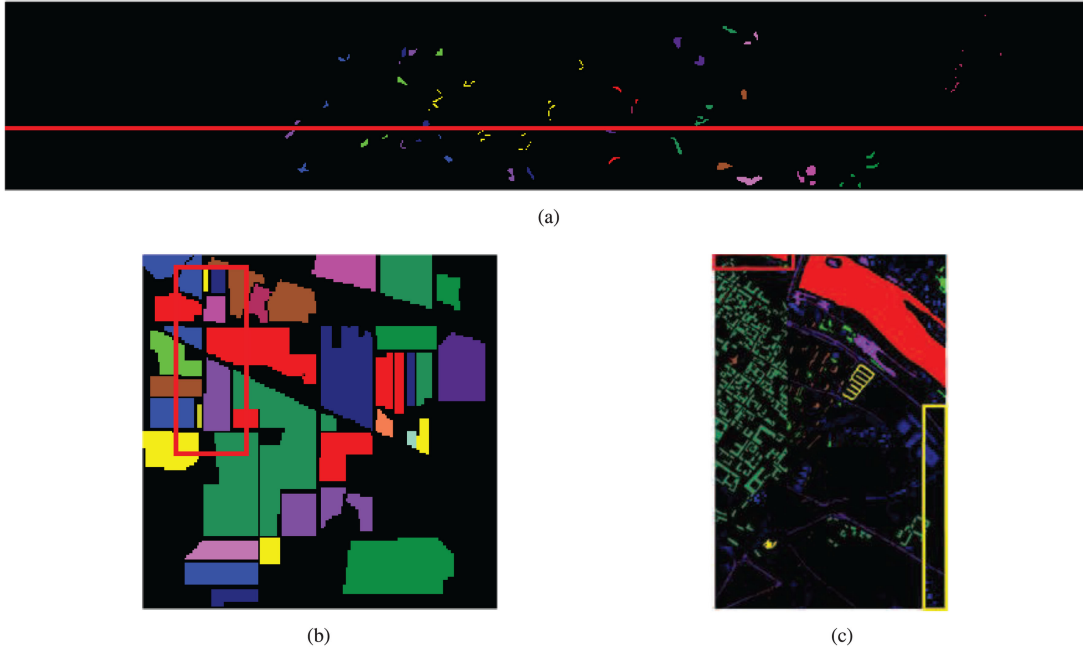


Fig. 3. Pseudocolor image of HSI dataset. (a) Botswana. (b) Indian Pines. (c) Pavia City.

---

**Algorithm 1: DABL.**


---

**Inputs:** MNF-based HSI representation  $\mathbf{X}$ , labeled samples of source domain  $\mathbf{Y}_s$ , domain adaptation parameter  $\alpha$ , manifold regularization parameter  $\beta$ , domain adaptation parameter  $\eta$ , manifold regularization parameter  $\gamma$ , feature dimensions of each group  $G^M$ , and number of nodes in MF per group  $d^M$ , number of nodes in EN  $d^E$ .

Step 1. Calculate the distribution importance parameter  $\mu$  according to (14).

Step 2. Calculate the weights  $\mathbf{A}_i$  according to (22).

Step 3. Calculate features in MF and EN from source and target domains  $\mathbf{Z}_s, \mathbf{Z}_t, \mathbf{H}_s, \mathbf{H}_t$  according to (29) and (30).

Step 4. Calculate the output-layer weights  $\mathbf{W}$  according to (33).

Step 5. Calculate the prediction labels  $\mathbf{Y}_t$  according to (35).

**Outputs:** Prediction labels  $\mathbf{Y}_t$ .

---

#### IV. EXPERIMENTS

##### A. HSI Datasets

To verify the validity and superiority of the proposed DABL, three real HSI datasets, including Botswana, Indian Pines, and Pavia City, are selected in our experiments.

Botswana dataset was acquired by researchers using the NASA EO-1 sensor over the Okavango Delta, Botswana. It consists of  $256 \times 1476$  pixels and 145 bands, including nine classes. The pseudocolor image of Botswana dataset is shown in Fig. 3(a), which is divided into two disjoint parts for domain

TABLE I  
NUMBER OF SAMPLES IN SOURCE AND TARGET DOMAINS SELECTED FROM BOTSWANA DATASET

Surface object	Source	Target	Total
F-grasses1	88	163	251
F-grasses2	136	79	215
Riparian	123	146	269
I-interior	126	77	203
A-woodlands	210	104	314
A-shrublands	58	43	101
Total samples	741	612	1353

adaptation with a red line. The region above the red line is considered as the source domain, and the region below the red line is considered as the target domain [41]. The selected two parts have similar land covers. For classification tasks, we selected six classes from both domains, i.e., F-grasses1, F-grasses2, Riparian, I-interior, A-woodlands, and A-shrublands, which are listed in Table I.

Indian Pines dataset was acquired by the ROSIS-03 sensor over the Indian Pines test site in North-Western Indiana, which consists of  $145 \times 145$  pixels and 224 bands. The pseudocolor image is shown in Fig. 3(b), the region in the red rectangle is regarded as the source domain, which contains the lines from 5 to 85 and columns from 10 to 40, and the others are treated as the target domain [29]. Both parts contain nine classes, i.e., Corn-notill, Corn-mintill, Corn, Grass-pasture, Grass-trees, S-notill, S-mintill, S-clean, and B-G-T-Driver, which are listed in Table II.

The third dataset is Pavia City. It was obtained by researchers using the ROSIS-03 sensor over Pavia, northern Italy. It consists of  $1096 \times 1096$  pixels and 102 bands, including nine classes.

TABLE II  
NUMBER OF SAMPLES IN SOURCE AND TARGET DOMAINS SELECTED FROM  
INDIAN PINES DATASET

Surface object	Source	Target	Total
Corn-notill	340	1088	1428
Corn-mintill	359	471	830
Corn	169	68	237
Grass-pasture	185	298	483
Grass-trees	270	460	730
S-notill	60	912	972
S-mintill	163	2292	2455
S-clean	198	395	593
B-G-T-Driver	89	297	386
Total samples	1833	6281	8114

TABLE III  
NUMBER OF SAMPLES IN SOURCE AND TARGET DOMAINS SELECTED FROM  
PAVIA CITY DATASET

Surface object	Source	Target	Total
Water	693	699	1392
Tree	133	1047	1180
Asphalt	185	142	327
Bitumen	35	24	59
Meadows	487	1005	1492
Total samples	1533	2917	4450

Because part of the data was discarded, the real dataset consists of  $1096 \times 715$  pixels and 102 bands. The pseudocolor image is shown in Fig. 3(c). The area in red rectangle, which contains the lines from 1 to 60 and columns from 1 to 225, is regarded as the source domain. The region in the yellow rectangle, which contains the lines from 380 to 1096 and columns from 620 to 715, is treated as the target domain [41]. Both regions contain five classes, i.e., Water, Tree, Asphalt, Bitumen, and Meadows. The details are given in Table III.

### B. Parameter Settings

According to the description of DABL, the adjustable parameters include: domain adaptation parameters  $\alpha$  and  $\eta$ , regularization parameters  $\beta$ ,  $\delta$ ,  $\lambda$ , and  $\gamma$ , feature dimensions of each group  $G^M$ , number of nodes in MF per group  $d^M$ , number of nodes in EN  $d^E$ , nearest neighbor parameter  $\psi$ , and ADMM parameters  $\rho$  and  $k$ . For single-domain classification problems, the whole dataset is generally divided into three sets: training set, testing set, and validation set. Based on the validation set, the cross-validation is commonly used to perform hyperparameter selection [42], [43]. But for the cross-domain classification problems, the source and target domains follow different distributions. Long *et al.* [18] stated that it is impossible to tune the optimal hyperparameters using cross-validation. Therefore, according to Long *et al.* [18], we use the empirically searching method to set the hyperparameters. The value ranges of the aforementioned parameters are  $\alpha, \eta, \beta \in \{0.01, 0.1, 1, 10, 100\}$ ,  $G^M \in \{5, 40, 75, 110, 145\}$ ,  $d^M \in \{5, 15, 25, 35, 45\}$ ,  $d^E \in \{250, 500, 750, 1000, 1250\}$ ,  $\psi \in \{1, 3, 5, 7, 9\}$ ,  $\lambda, \rho \in \{0.001, 0.01, 0.1, 1, 10\}$ ,  $\delta \in \{2^{-40}, 2^{-30}, 2^{-20}, 2^{-10}, 1\}$ , and  $k \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ . Given  $\gamma = 0.01$ , the relationships

between these parameters and the overall accuracy (OA) are shown in Figs. 4–9. Following can be observed from Figs. 4–9.

- 1) On the one hand, small  $G^M$  and  $d^M$  indicate that the dimension of features in MF is low. The input HSI data cannot be adequately represented. On the other hand, large  $G^M$  may lead to feature redundancy in MF. Similarly, if  $d^E$  is too small, broad expansion cannot be achieved sufficiently, whereas large  $d^E$  and  $d^M$  may lead to feature redundancy in EN. Therefore, for Bostwana dataset, we set  $G^M = 110$  and  $d^M = 35$ . For Indian Pines dataset, we set  $G^M = 20$  and  $d^M = 23$ . For Pavia dataset, we set  $G^M = 130$  and  $d^M = 10$ . For the three datasets above, the number of EN layer nodes is set to be  $d^E = 1000$ .
- 2) Larger  $\alpha$  and  $\eta$  means that the domain adaptation part of the DABL plays a more important role, which is suitable to the case where the source and target domains are quite different. Choosing a suitable manifold regularization parameter  $\beta$  and  $\gamma$  can sufficiently represent the complex manifold structure of HSI data. Therefore, for Bostwana dataset, we set  $\alpha = \eta = 0.1$  and  $\beta = 10$ . For Indian Pines dataset, we set  $\alpha = \eta = 0.1$  and  $\beta = 10$ . For Pavia City dataset, we set  $\alpha = \eta = 0.1$  and  $\beta = 100$ .
- 3) On the one hand, small  $\psi$  may lead to misclassification. On the other hand, large  $\psi$  may increase the amount of calculation. Therefore, we set  $\psi = 3$ . Large  $\lambda$  may lead to overfitting. On the contrary, small  $\lambda$  may result in under fitting. Therefore, we set  $\lambda = 1$ . In addition, according to the parameter settings of BLS and ADMM in [30] and [31], respectively, we set  $\delta = 2^{-30}$ ,  $\rho = 1$ , and  $k = 10^{-3}$ .

### C. Comparative Experiments

To demonstrate the classification performance of the proposed DABL, the following nine methods are selected for comparison.

- 1) Traditional classification method: SVM [8].
- 2) Transfer learning methods: TCA [44], JDA [18], DST-ELM [19], and manifold embedded distribution alignment [39].
- 3) Deep domain adaptation method: DANN [20].
- 4) Broad learning methods: DABL without manifold regularization and domain adaption, i.e., BLS [30] and DABL without manifold regularization (DABL1), and DABL with the following hyperparameters (DABL2):  $G^M = 110$ ,  $d^M = 20$ ,  $d^E = 1000$ ,  $\alpha = \eta = 0.1$ ,  $\beta = 100$ ,  $\psi = 3$ ,  $\lambda = 1$ ,  $\delta = 2^{-30}$ ,  $\rho = 1$ , and  $k = 10^{-3}$ .

All methods were implemented in MATLAB 2017a on an Intel i5-6500 CPU with 8-GB memory. To ensure fair comparison, inputs of each aforementioned method were preprocessed with MNF and the optimal parameters of seven methods were selected with cross-validation. Each experiment was repeated ten times to get the average value to reduce the effects of random factors. Five evaluating indexes are considered: the per-class accuracy, the average accuracy (AA), the overall accuracy (OA), the Kappa coefficient, and the consumed time. The reported consumed time here means the training and testing time of classifier. OA is defined by the ratio between the number of correctly classified pixels to the total number of pixels in the



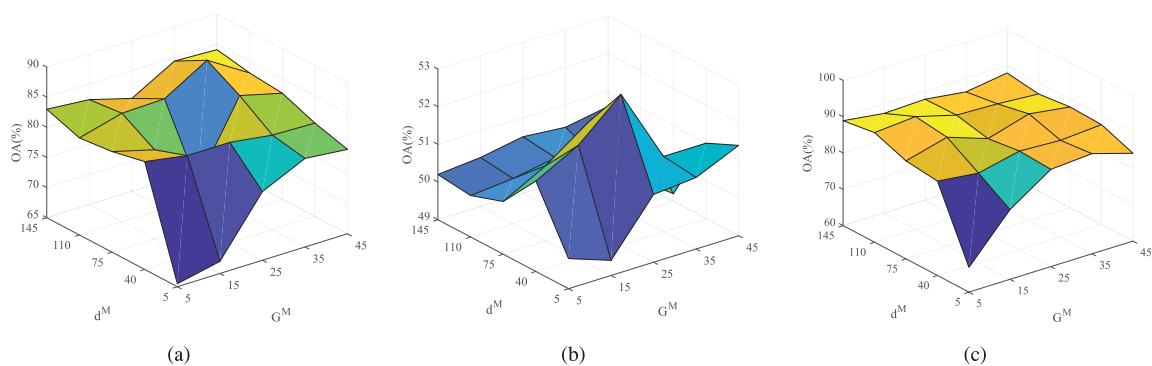


Fig. 4. Variation of OA over parameters  $G^M$  and  $d^M$  on different HSI datasets. (a) Botswana. (b) Indian Pines. (c) Pavia City.

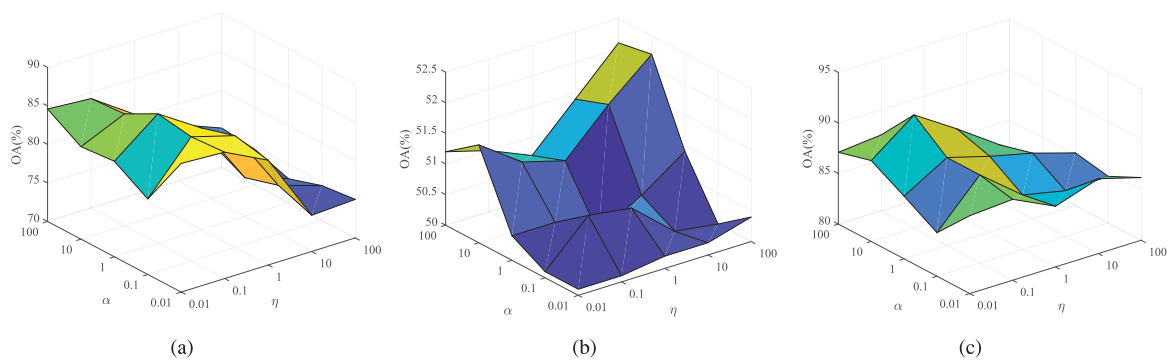


Fig. 5. Variation of OA over parameters  $\alpha$  and  $\eta$  on different HSI datasets. (a) Botswana. (b) Indian Pines. (c) Pavia City.

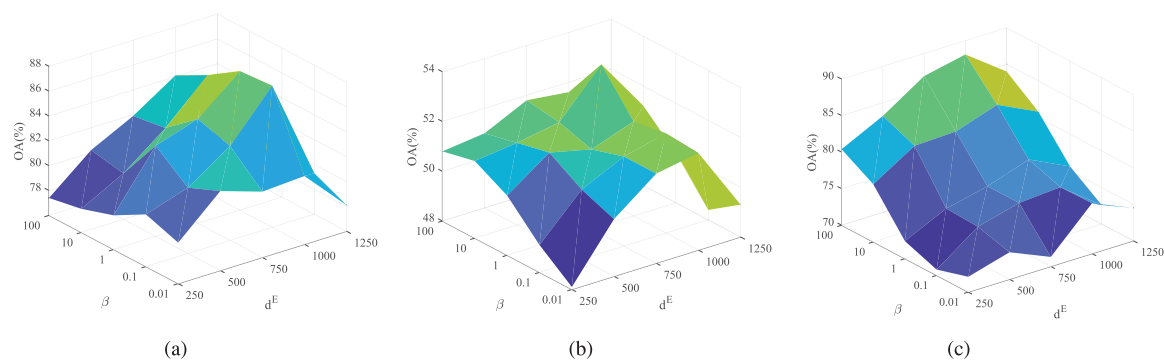


Fig. 6. Variation of OA over parameters  $\beta$  and  $d^E$  on different HSI datasets. (a) Botswana. (b) Indian Pines. (c) Pavia City.

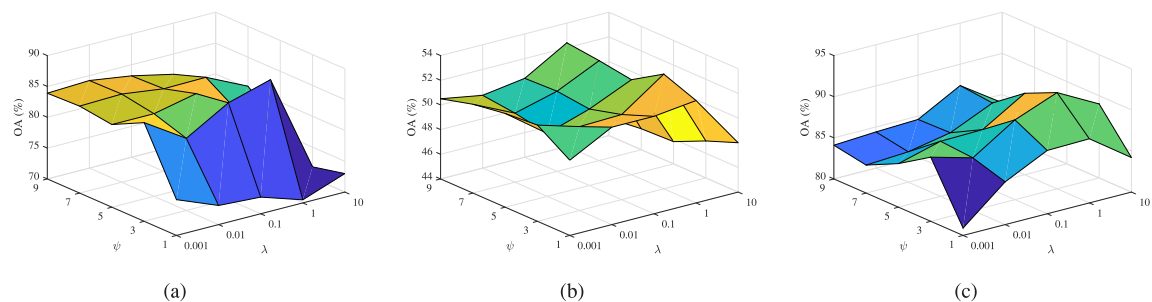


Fig. 7. Variation of OA over parameters  $\lambda$  and  $\psi$  on different HSI datasets. (a) Botswana. (b) Indian Pines. (c) Pavia City.

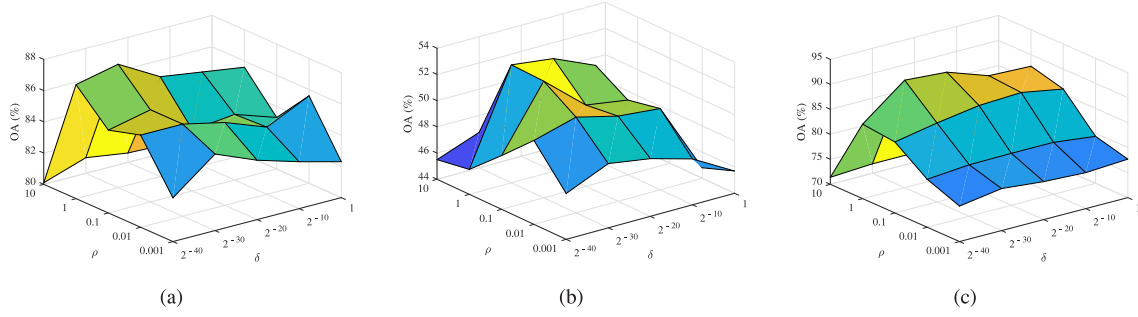


Fig. 8. Variation of OA over parameters  $\delta$  and  $\rho$  on different HSI datasets. (a) Botswana. (b) Indian Pines. (c) Pavia City.

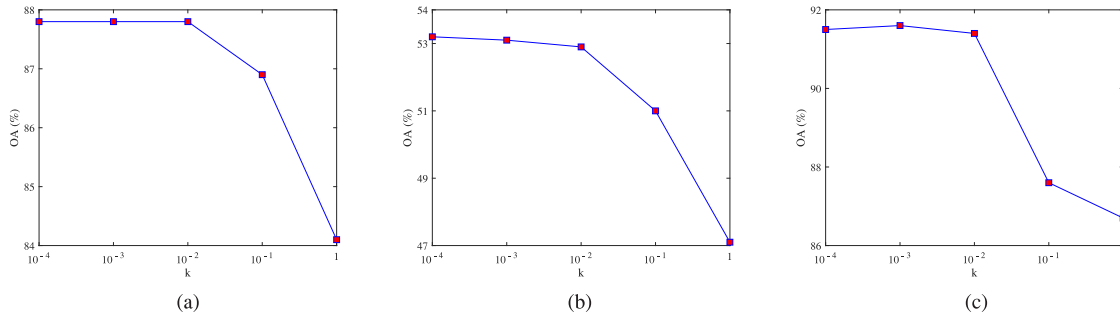


Fig. 9. Variation of OA over parameter  $k$  on different HSI datasets. (a) Botswana. (b) Indian Pines. (c) Pavia City.

TABLE IV  
COMPARISON OF CLASSIFICATION PERFORMANCE ON BOTSWANA

Surface object	SVM [8]	TCA [44]	JDA [18]	DST-ELM [19]	EMDA [39]	DANN [20]	BLS [30]	DABL1	DABL2	DABL
F-grasses1 (%)	<b>100</b>	<b>100</b>	<b>100</b>	93.02	<b>100</b>	<b>100</b>	<b>100</b>	97.67	<b>100</b>	<b>100</b>
F-grasses2 (%)	<b>99.23</b>	68.10	69.94	57.06	62.58	78.53	99.07	74.23	98.77	95.99
Riparian (%)	95.18	98.73	97.47	<b>100</b>	<b>100</b>	91.14	61.72	98.73	<b>100</b>	<b>100</b>
I-interior (%)	86.56	89.04	<b>89.04</b>	76.71	88.56	71.23	70.27	72.60	67.81	70.47
A-woodlands (%)	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	81.82	96.20	89.62	98.70	94.31
A-shrublands (%)	36.80	74.04	75.00	74.04	79.81	71.15	77.33	80.01	73.08	<b>80.04</b>
AA (%)	86.30	88.32	88.58	88.47	88.49	82.31	84.10	85.48	89.72	<b>90.14</b>
OA (%)	71.40	84.31	84.80	79.10	84.83	79.19	79.08	81.69	87.25	<b>87.77</b>
Kappa	0.6497	0.8077	0.8136	0.7431	0.8106	0.7429	0.7426	0.7751	0.8425	<b>0.8760</b>
Time (s)	2.15	1.23	4.00	20.81	11.44	1614.04	<b>0.22</b>	5.89	28.02	11.42

The bold entities represent the optimal values of indexes.

testing set. AA refers to the average of accuracies in all classes, and Kappa coefficient is the percentage of agreement corrected by the number of agreements that would be expected purely by chance. The classification results on three HSI datasets are listed in Tables IV–VI.

Following can be observed from Tables IV–VI.

- 1) Among the three HSI datasets, all ten methods have the lowest OAs and Kappa coefficients on Indian Pines. This is due to the high similarity between classes in the Indian Pines dataset. For instance, the corn-notill, corn-mintill, and corn belong to the same class in essence. Because of the similar spectral features between classes, there is a high degree of mixture in the feature space distribution.
- 2) Compared with TCA, JDA can obtain higher OAs and Kappa coefficients. The reason is that JDA further adapts the conditional distributions between two domains, which can enhance the model ability of discriminating target data.

- 3) SVM and BLS are both nontransfer learning methods. Compared with SVM, BLS not only has higher OAs and Kappa coefficients, but also consumes shorter time. There are mainly three reasons. First, the SAE is used to extract the features of original HSI, so that better feature representation of original HSI can be obtained. Second, BLS can map MF with random weights to achieve nonlinear broad expansion and feature enhancement, so that the overall BLS has a strong function approximation ability. Finally, the weights of MF to EN in BLS are generated randomly instead of a complicated training process, and the output layer weight can be easily obtained with the ridge regression theory. Therefore, BLS is efficient.
- 4) It can be seen from Table VI that DANN obtains the highest AA of 94.54%. The reason for this phenomenon is because DANN classifies the minority classes better, whereas DABL is better in the majority classes tree and meadows. However, DABL achieves higher OA and

TABLE V  
COMPARISON OF CLASSIFICATION PERFORMANCE ON INDIAN PINE

Surface object	SVM [8]	TCA [44]	JDA [18]	DST-ELM [19]	EMDA [39]	DANN [20]	BLS [30]	DABL1	DABL2	DABL
Corn-notill (%)	53.86	52.76	62.41	58.45	64.34	<b>78.21</b>	56.23	56.34	49.63	59.16
Corn-mintill (%)	24.42	40.98	43.10	41.77	40.34	44.46	<b>45.12</b>	43.10	18.47	30.76
Corn (%)	13.24	<b>38.24</b>	<b>38.24</b>	38.48	27.94	19.12	13.77	16.47	4.41	21.01
Grass-pasture (%)	53.36	51.34	50.67	53.70	50.34	53.36	49.55	50.34	55.00	<b>55.17</b>
Grass-trees (%)	96.30	96.52	95.65	97.77	98.48	77.84	95.00	<b>98.17</b>	83.04	85.94
S-notill (%)	1.32	0.11	1.64	0.35	3.18	0.66	0.87	1.64	<b>6.90</b>	<b>6.90</b>
S-mintill (%)	54.32	55.15	50.35	<b>64.19</b>	47.25	47.11	63.15	57.15	69.24	60.11
S-clean (%)	<b>98.48</b>	34.94	36.71	44.91	64.81	70.89	49.87	81.77	79.95	82.17
B-G-T-Driver (%)	57.58	59.60	70.03	62.31	84.85	60.55	78.44	82.15	75.08	<b>86.21</b>
AA (%)	50.32	47.74	49.87	51.33	53.50	50.24	50.22	54.13	49.08	<b>54.16</b>
OA (%)	49.82	47.44	48.10	52.62	49.86	52.84	52.59	52.66	52.85	<b>53.74</b>
Kappa	0.3772	0.3488	0.3595	0.3995	0.3834	0.4012	0.4021	0.4041	0.3929	<b>0.4116</b>
Time (s)	1.56	45.61	88.35	160.64	139.15	3417.5	<b>0.46</b>	101.44	103.46	103.16

The bold entities represent the optimal values of indexes.

TABLE VI  
COMPARISON OF CLASSIFICATION PERFORMANCE ON PAVIA CITY

Surface object	SVM [8]	TCA [44]	JDA [18]	DST-ELM [19]	EMDA [39]	DANN [20]	BLS [30]	DABL1	DABL2	DABL
Water (%)	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Tree (%)	57.02	56.54	50.91	46.80	67.15	81.57	69.34	71.17	77.75	<b>82.68</b>
Asphalt (%)	48.59	30.99	76.76	44.67	87.32	<b>99.30</b>	78.87	78.80	53.11	60.57
Bitumen (%)	25.00	4.17	4.17	66.67	<b>100</b>	99.7	50.00	50.00	91.67	87.50
Meadows (%)	99.90	99.40	99.80	99.30	<b>100</b>	92.14	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
AA (%)	66.10	58.22	66.33	71.49	90.89	<b>94.54</b>	79.64	79.99	84.51	86.15
OA (%)	81.41	80.04	80.39	78.65	87.58	90.64	87.56	88.21	89.61	<b>91.75</b>
Kappa	0.7462	0.7253	0.7351	0.7075	0.8411	0.8700	0.8377	0.8382	0.8631	<b>0.8979</b>
Time (s)	4.47	12.69	27.43	117.57	43.93	5459.11	<b>1.39</b>	13.47	30.41	14.41

The bold entities represent the optimal values of indexes.

Kappa coefficient than DANN. It is known that for class-imbalance classification problems, AA does not reflect the performance of classifier very well, whereas OA is more objective. Therefore, in the field of HSI classification, compared with the per-class accuracy and AA, OA and Kappa coefficient are more important indexes [45]. In addition, the consumed time of DANN is 5459.11 s, which is almost 380 times of DABL. Therefore, we can conclude that DABL outperforms DANN.

- 5) DANN achieves the second high OAs and Kappa coefficients on Indian Pines and Pavia City datasets, but it consumes the longest time among the ten methods, which is not suitable for the situation requiring high real time. In addition, DANN achieves a low accuracy on the Botswana dataset, which is caused by insufficient training samples.
- 6) BLS has achieved high classification accuracy on the Indian Pines and Pavia City datasets, even surpassing some transfer learning methods. The reason is that the nonlinear mapping from MF to EN in BLS achieves the broad expansion of MF and enhances the generalization ability of BLS.
- 7) Among the ten methods, DABL obtains the highest OAs and Kappa coefficients on all three HSI datasets. The main reasons are discussed as following. First, DABL makes full use of the strong function approximation capability of BLS to achieve more accurate mapping from feature space to class space. Second, by adding manifold regularization and domain adaptation terms to SAE, the features learned in MF not only maintain the manifold structure but also enhance the domain invariance. Finally, the domain

adaptation regularization term is also added into the output layer of DABL to achieve the classifier adaptation. However, it should be noted that the classification performance of DABL is sensitive to the setting of hyperparameters.

## V. CONCLUSION

An HSI classification method, named DABL, is proposed in this article. First, to reduce the distribution difference and maintain manifold structure, the proposed DABL adapts both distributions between source and target domains and adds the manifold regularization term. Then, by mapping the MF to EN with randomly generated weights, the features achieve broad expansion. Furthermore, by combining the domain adaptation and manifold regularization terms in the objective function, we further reduce the distribution difference and maintain manifold structure. Finally, the objective function can be easily obtained with the ridge regression theory. Experimental results on three real HSI datasets demonstrate the proposed DABL can obtain higher classification accuracy than several methods.

## REFERENCES

- [1] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.
- [2] D. Fasbender *et al.*, "Evaluating NDVI data continuity between SPOT-VEGETATION and PROBA-V missions for operational yield forecasting in North African countries," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 795–804, Feb. 2016.
- [3] 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.

- [4] K. G. Elham, M. S. Helfroush, and H. Danyali, "Sparse-based classification of hyperspectral images using extended hidden Markov random fields," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4101–4112, Nov. 2018.
- [5] B. Tu, J. Wang, G. Zhang, X. Zhang, and W. He, "Texture pattern separation for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3602–3614, Sep. 2019.
- [6] X. D. Kang, C. C. Li, S. T. Li, and H. Lin, "Classification of hyperspectral images by Gabor filtering based deep network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1166–1178, Apr. 2018.
- [7] J. Xia, L. Bombrun, Y. Berthoumieu, C. Germain, and P. Du, "Spectral spatial rotation forest for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 10, pp. 4605–4613, Oct. 2017.
- [8] F. Melgani and L. Bruzzone, "Support vector machines for classification of hyperspectral remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
- [9] K. Tan, J. Zhang, Q. Du, and X. Wang, "GPU parallel implementation of support vector machines for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 10, pp. 4647–4656, Oct. 2015.
- [10] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.
- [11] M. S. Aydemir and G. Bligin, "Semisupervised hyperspectral image classification using deep features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3615–3622, Sep. 2019.
- [12] Y. Ding, S. Pan, and Y. Chong, "Robust spatial-spectral block-diagonal structure representation with fuzzy class probability for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1747–1762, Mar. 2020.
- [13] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.
- [14] J. Munoz-Mari, D. Tuia, and G. Camps-Valls, "Semi-supervised classification of remote sensing images with active queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3751–3763, Oct. 2012.
- [15] 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.
- [16] 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.
- [17] X. Zhou and S. Prasad, "Domain adaptation for robust classification of disparate hyperspectral images," *IEEE Trans. Comput. Imag.*, vol. 3, no. 4, pp. 822–826, Dec. 2017.
- [18] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2200–2207.
- [19] Y. Chen, S. Song, L. Yang, and C. Wu, "Domain space transfer extreme learning machine for domain adaptation," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1909–1922, May 2019.
- [20] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by back-propagation," in *Proc. 32th Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
- [21] Z. Sun, C. Wang, H. Wang, and J. Li, "Learn multiple-kernel SVMs for domain adaptation in hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1224–1228, Sep. 2013.
- [22] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample problem," in *Proc. Conf. Neural Inf. Process. Syst.*, May 2007, pp. 513–520.
- [23] D. Tuia, D. Marcos, and G. Camps-Valls, "Multi-temporal and multi-source remote sensing image classification by nonlinear relative normalization," *ISPRS J. Photogrammetry Remote Sens.*, vol. 120, pp. 1–12, Oct. 2016.
- [24] J. S. Xia, N. Yokoya, and A. Iwasaki, "Ensemble of transfer component analysis for domain adaptation in hyperspectral remote sensing image classification," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2017, pp. 4762–4765.
- [25] H. Sun, S. Liu, S. Zhou, and H. Zhou, "Transfer sparse subspace analysis for unsupervised cross-view scene model adaptation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 7, pp. 2901–2909, Jul. 2016.
- [26] X. Li, L. Liang, 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. 5863–5872, Jun. 2019.
- [27] E. Riz, B. Demir, and L. Bruzzone, "Domain adaptation based on deep denoising auto-encoders for classification of remote sensing images," in *Proc. Image Signal Process. Remote Sens.*, 2016, vol. 10004, Art. no. 100040K.
- [28] 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.
- [29] 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.
- [30] C. L. P. Chen and Z. Liu, "Broad learning system: An effective and efficient incremental learning system without the need for deep architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 10–24, Jul. 2017.
- [31] S. Feng and C. L. P. Chen, "Fuzzy broad learning system: A novel neuro-fuzzy model for regression and classification," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 414–424, Feb. 2020.
- [32] J. Jin, Z. Liu, and C. L. P. Chen, "Discriminative graph regularized broad learning system for image recognition," *Sci. China-Inf. Sci.*, vol. 61, no. 11, Nov. 2018, Art. no. 112209.
- [33] Y. Kong, X. Wang, Y. Cheng, and C. L. P. Chen, "Hyperspectral imagery classification based on semi-supervised broad learning system," *Remote Sens.*, vol. 10, no. 5, Apr. 2018, Art. no. 685.
- [34] Y. Kong, Y. Cheng, C. L. P. Chen, and X. Wang, "Hyperspectral imagery clustering based on unsupervised broad learning system," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1741–1745, Nov. 2019.
- [35] E. Othman, Y. Bazi, F. Melgani, H. Alhichri, N. Alajlan, and M. Zuair, "Domain adaptation network for cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4441–4456, Aug. 2017.
- [36] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.
- [37] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, "Semi-supervised transfer component analysis for domain adaptation in remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3550–3564, Jul. 2015.
- [38] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye, "A two-stage weighting framework for multi-source domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2011, pp. 505–513.
- [39] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *Proc. ACM Multimedia Conf.*, 2018, pp. 402–410.
- [40] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2006, pp. 137–144.
- [41] 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.
- [42] S. Kaufman, S. Rosset, and C. Perlich, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 4, pp. 1–21, Dec. 2012.
- [43] A. Santara *et al.*, "BASS Net: Band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5293–5301, Sep. 2017.
- [44] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
- [45] R. Hang, Q. Liu, H. Song, and Y. Sun, "Matrix-based discriminant subspace ensemble for hyperspectral image spatial-spectral feature fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 783–794, Feb. 2016.



**Haoyu Wang** received the M.S. degree, in 2017 from the China University of Mining and Technology, Xuzhou, China, where he is currently working toward the Ph.D. degree with the School of Information and Control Engineering.

His main research interests include hyperspectral image analysis and transfer learning.





**Xuesong Wang** (Member, IEEE) received the Ph.D. degree in control science and technology from the China University of Mining and Technology, Xuzhou, China, in 2002.

She is currently the Dean of the Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, the Dean of Xuzhou Key Laboratory of Artificial Intelligence and Big Data, and a Professor with the School of Information and Control Engineering, China University of Mining and Technology. Her main research interests include

machine learning, bioinformatics, and artificial intelligence.

Prof. Wang is currently an Associate Editor for the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS and *International Journal of Machine Learning and Cybernetics*.



**Yuhu Cheng** (Member, IEEE) received the Ph.D. degree in control science and technology from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently a Professor with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China. His main research interests include machine learning and intelligent system.



**C. L. Philip Chen** (Fellow, IEEE) received the M.S. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988.

He is currently the Dean of the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, and a Chair Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China. His current research

interests include computational intelligence, systems, and cybernetics.

Dr. Chen is a Fellow of AAAS, the Chinese Association of Automation, and HKIE. He is also the Chair of TC 9.1 Economic and Business Systems of International Federation of Automatic Control. He is also an Accreditation Board of Engineering and Technology Education Program Evaluator for Computer Engineering, Electrical Engineering, and Software Engineering programs. He has been the Editor-in-Chief for the IEEE TRANSACTIONS ON CYBERNETICS since 2020 and an Associate Editor for several IEEE Transactions. He was the President of the IEEE Systems, Man, and Cybernetics Society from 2012 to 2013.