Spectral–Spatial Weighted Kernel Manifold Embedded Distribution Alignment for Remote Sensing Image Classification

Yanni Dong, *Member, IEEE*, Tianyang Liang, Yuxiang Zhang, and Bo Du, *Senior Member, IEEE*

***Abstract*—Feature distortions of data are a typical problem in remote sensing image classification, especially in the area of transfer learning. In addition, many transfer learning-based methods only focus on spectral information and fail to utilize spatial information of remote sensing images. To tackle these problems, we propose spectral–spatial weighted kernel manifold embedded distribution alignment (SSWK-MEDA) for remote sensing image classification. The proposed method applies a novel spatial information filter to effectively use similarity between nearby sample pixels and avoid the influence of nonsample pixels. Then, a complex kernel combining spatial kernel and spectral kernel with different weights is constructed to adaptively balance the relative importance of spectral and spatial information of the remote sensing image. Finally, we utilize the geometric structure of features in manifold space to solve the problem of feature distortions of remote sensing data in transfer learning scenarios. SSWK-MEDA provides a novel approach for the combination of transfer learning and remote sensing image characteristics. Extensive experiments have demonstrated that the proposed method is more effective than several state-of-the-art methods.**

Manuscript received February 6, 2020; revised May 29, 2020; accepted June 19, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61801444, Grant 61701452, Grant 62041105, and Grant 6182211; in part by the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) under Grant CUG170687; in part by the Open Research Fund of Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences under Grant 2018LDE004; in part by the Open Research Project of the Hubei Key Laboratory of Intelligent Geo-Information Processing under Grant KLIGIP-2018A01; in part by the Open Research Fund of CAS Key Laboratory of Spectral Imaging Technology under Grant LSIT201921W; in part by the State Key Laboratory of Integrated Services Networks (Xidian University) under Grant ISN20-07; in part by the China Postdoctoral Science Fund under Grant 2017M612533; in part by the China Scholarship Council under Grant 201806415014; and in part by the Hong Kong Scholars Program under Grant XJ2018012. This article was recommended by Associate Editor J. Han. *(Corresponding author: Yuxiang Zhang.)*

Yanni Dong is with the Hubei Subsurface Multi-Scale Imaging Key Laboratory, Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China, also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China, and also with the Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong (e-mail: dongyanni@cug.edu.cn).

Tianyang Liang and Yuxiang Zhang are with the Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China (e-mail: liangtianyang@cug.edu.cn; zhangyx@cug.edu.cn).

Bo Du is with the School of Computer Science, Wuhan University, Wuhan 430079, China (e-mail: gunspace@163.com).

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

Digital Object Identifier 10.1109/TCYB.2020.3004263

***Index Terms*—Classification, Grassmann manifold, remote sensing, spatial and spectral information, transfer learning, weighted kernel.**

I. Introduction

IMAGE classification is a key research direction in the field of remote sensing; it aims to classify every pixel as a certain surface object [1]–[3]. It can benefit from the labeled pixels. Therefore, higher classification accuracy can be obtained when more labeled pixels are available [4], [5]. Since labeling a remote sensing image is very laborious by artificial operation, facing massive remote sensing data, scholars seek to find an automatic and efficient approach to get this job done. An intuitive idea is using the trained classifier on the new image to be processed. However, a remote sensing scene is usually mosaiced by a series of images obtained under different situations, such as cloud and light. Therefore, the classifier trained by one of these images cannot be used for other images directly. Similar things happen in researching the study of the changes in an area by two images from different time, and the difference may be larger when they are produced by different sensors.

To solve this issue, we consider a transfer learning setting in which the knowledge from source domain can be transferred into the target domain to obtain predictive labels [6], [7]. Transfer learning is a powerful route to classify multitemporal or multisource remote sensing images.

As such, transfer learning or domain adaptation (a special form of transfer learning), has received much attention in the research community [8]–[10]. Various transfer learning methods have been presented and divided into three categories, namely: 1) instance-based methods; 2) classifier-based methods; and 3) feature-based methods [11]–[13].

Instance-based methods usually set more weights to important samples. Samples with poor performance usually reduce their weights by a different criterion or even remove them [14], [15]. Jiang and Zhai [16] attempted to remove samples with poor performance from the source domain and increased the weights of labeled samples from the target domain in accordance with the condition distribution of both domains. Dai *et al.* [17] put forward a framework called TrAdaBoost, which initially trained a basic classifier and then reweighted source domain samples by the classification

error in each iteration. Persello [18] reweighted each source domain sample by its mean cosine-angle similarity with the target domain samples of the same class. Long *et al.* [19], Yang *et al.* [20], and Peng *et al.*[21] added the $l1$ and $l2$ norm terms on the transform matrix of the source domain to reweight them. Shi *et al*. [22] reconstructed each sample by using probability distribution with other neighbor samples. Li *et al*. [23] presented a framework called iterative reweighting heterogeneous transfer learning (IRHTL). This approach is used to reweight source domain samples by their distance from the separating hyperplane.

Classifier-based methods generally adapt the classifier trained in the source domain into the target domain [24], [25]. The most commonly used classifier is the support vector machine (SVM). Bruzzone and Prieto [26] presented a maximum-likelihood classifier, which was trained by the source domain and can update parameters to adapt the target domain. They introduced another work in [27], which initially trained an SVM classifier on the source domain. Then, they iteratively added target domain samples with high confidence and removed samples from the source domain to increase the adaptive ability of the classifier. Yang *et al*. [28] put forward adaptive SVM (A-SVM) to adapt the margin bound of SVM from the source domain into the target domain, while maximizing the margin bound between classes. Guo *et al*. [29] extended ASVM into temporal A-SVM (TASVM), which allowed to position the SVM of each domain in any place. In their subsequent article [30] for multitemporal remote sensing images, they initialized an SVM classifier on an image and added new image by time series in turn, and fine-tuned the classifier with TASVM. Xu *et al*. [31] proposed a method called domain adaptation with parameter transfer, which aligned the parameters of the target domain classifier to source domain by transferring the parameters of the extreme learning machine. On the basis of the classifier trained in the source domain, Rajan *et al*. [32] developed a new classifier by using the binary hierarchical classifier to obtain the hierarchical class information of images in other regions or phases.

Feature-based methods can also be divided into three kinds: 1) aligning domains by using their distributions; 2) finding a common feature space; or 3) selecting a feature subset from all features [33]–[35]. The following research is about these three kinds of methods. For the former kind, maximum mean discrepancy (MMD) is the most widely used distance function. Pan *et al.* [36] and Long *et al.* [37] presented transfer component analysis (TCA) and joint distribution adaptation (JDA), respectively, which used the kernel function to solve the MMD of marginal and condition distributions. Then, Wang *et al*. [38] introduced a method called manifold embedded distribution alignment (MEDA), which extended JDA into Grassmann manifold space [39] with the structural risk minimization (SRM) and the Laplacian regularization (LR). Different from these works, Tuia *et al*. [40] linked centroid alignment and graph matching, and Ma *et al*. [41] considered overall centroid alignment with class centroid and covariance alignment. For the second kind, Frenando *et al*. [42] put forward subspace alignment (SA), which constructed a transform matrix between the principal component analysis (PCA) subspaces of both domains. Based on this, Sun *et al*. [43] replaced the PCA subspace of the target domain with the subspace of partial least squares correlation. In another article, Sun *et al.* [44] attempted to learn a transform matrix to project source domain and target domain into a common sparse space by minimizing MMD and preserving the original data variance. On the contrary, Gong *et al*. [45] viewed domain adaptation as an integral of geometric flow between two points representing the PCA subspace of source and target domains in Grassmann manifold. In addition, a simple but powerful method called correlation alignment (CORAL) was proposed by Sun *et al.* [46], where the covariance between the target domain and transferred source domain is aligned. For the last kind, Persello and Bruzzone [47] used the kernel method to measure the correlation between feature and classifier and domain invariance; then, feature bands were selected.

However, most of the above methods implement transfer learning in the original space, where feature distortions may occur, or only use a single kernel approach, probably leading to undesirable performance. Thus, removing feature distortions and achieving better performance is a challenging problem for remote sensing image. Not all pixels of a remote sensing image can be regarded as samples. In addition, these sample pixels have some connections of geometric shape or texture. Nonetheless, many transfer learning methods only focus on spectral information and fail to use spatial information of remote sensing image. Therefore, utilizing these connections is another challenging problem for the remote sensing image.

To alleviate these problems, we propose a method called spectral–spatial weighted kernel MEDA (SSWK-MEDA). Since the adjacent pixels in space are more likely to have the same properties and labels, we apply a fixed window size for every sample pixel and calculate the mean properties of all sample pixels in these windows, except for those in the center of the windows. The results obtained are called background sample pixels. Then, we balance the kernel weights between background sample pixels and original sample pixels by the average relevance of each pair of corresponding sample pixels.

Our main contributions are summarized as follows.

1) A novel spatial information filter is proposed for remote sensing image classification, which can effectively use the similarity of nearby sample pixels and avoid the influence of nonsample pixels.
2) We construct a complex kernel by spatial kernel and spectral kernel with different weights, which are determined by the measurement of their relative importance.
3) We introduce Grassmann manifold embedding, which can solve the problem of feature distortions of remote sensing data in transfer learning scenarios.

In general, a complex kernel by spatial kernel and spectral kernel with different weights is constructed to adaptively balance the importance between spectral and spatial information of the remote sensing image. Then, we utilize the geometric structure of features in manifold space to solve the problem of feature distortions of remote sensing data in transfer learning scenarios. SSWK-MEDA provides a novel approach for the

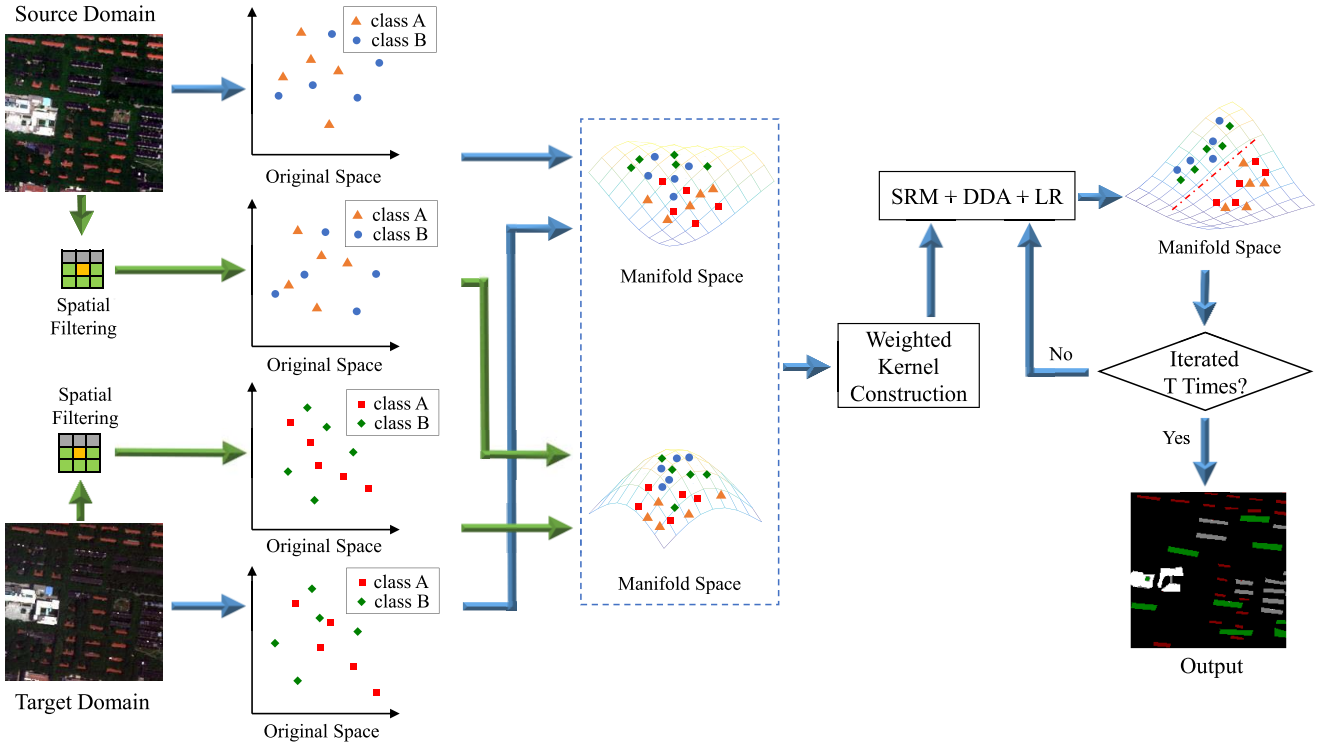

Fig. 1. Flowchart of the proposed SSWK-MEDA method.

combination of transfer learning and remote sensing image characteristics.

The remainder of this article is organized as follows. Section II presents the steps of the proposed method. Section III describes the proposed method. Section IV verifies the proposed method by a number of experiments and discussions. Section V provides a summary of the study and directions of future research.

II. Related Work

Filter Window: Spatial overall centroid alignment, class centroid, and covariance alignment [41] applied a filter window to extract the mean spectrum of all sample pixels in it to achieve more reliable spectral features and replace the original spectrum. Different from it, we only calculate the mean spectrum of the surrounding sample pixels for each sample pixel, called spatial features. It can prevent the interference of pixels, which do not belong to the samples. We still place the original spectrum of sample pixels in an important position, combining them with the spectrum of the surrounding sample pixels in subsequent operations.

Composite Kernel: The deep adaptation network [48] was proposed to construct a compound kernel by some weighted kernels. On this basis, we present a different complex kernel for remote sensing. This kernel is combined by the kernel of the original spectral features and the kernel of filtered spatial features. It can adaptively weight the relative importance between the two component kernels by their mean correlation coefficients.

Distribution Alignment: The distribution between the source domain and the target domain may be very different. Thus, many methods have been proposed to minimize the discrepancy. MMD was first innovated in TCA for adapting marginal distribution. Thereafter, JDA attempted to adapt marginal and conditional distributions using a similar method as TCA. And MEDA measured the importance between marginal distribution and conditional distribution and avoided the feature distortion of TCA and JDA by manifold embedding. However, they were not designed for remote sensing image processing. Thus, they cannot utilize the spatial information of remote sensing images. SSWK-MEDA uses spatial filters to obtain spatial information and combine it with its original spectral information in a weighted kernel approach.

III. Methodology

Given an unlabeled domain $\mathbf{X}_t = \{x_{t_i}\}_{i=1}^{m}$ with m samples, classifying its samples is difficult because no labeled information is available. A similar but different domain $\mathbf{X}_s = \{x_{s_i}, Y_{s_i}\}_{i=m+1}^{m+n}$ with n samples exists with the same types of labeled information Y with $\mathbf{X}_t$. Our goal is to find a classifier f, which can predict the labels of the target domain $\mathbf{X}_t$ by using the knowledge from the source domain $\mathbf{X}_s$.

The flowchart of the proposed SSWK-MEDA method is shown in Fig. 1. Source and target domain images initially go through spatial filters to obtain their spatial features. Subsequently, spectral and spatial features are projected into manifold space independently. Then, a composite kernel is constructed by adaptive weighting the kernels of spectral and spatial manifold features. Finally, a classifier can be learned by summarizing SRM, dynamic distribution alignment (DDA), and LR, which utilizes geometric properties

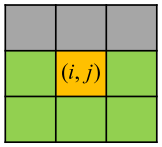

Fig. 2. Example of pixels.

among data samples. To obtain relatively high accuracy, the classifier can be iterated T times.

A. Spatial Filtration

For a remote sensing image $\mathbf{I}^{spe}$ with l bands and C classes, closer pixels are usually more similar. However, different classes of pixels may be very close to each other. In this case, we apply a $k \times k$ size window filter for all the sample pixels in $\mathbf{I}^{spe}$. The mean spectrum value of every adjacent sample pixel in the window is calculated for each sample pixel in each band. These pixels in the window are assumed to have the same class as the central pixel. The result of filtering is called $\mathbf{I}^{spa}$. Fig. 2 shows an example of sample pixels. Each cell in it is a sample pixel. The yellow cell with a coordinate (i, j) is the currently positioned and labeled sample pixels. The green cells are adjacent sample pixels, and the gray cells are nonsample adjacent pixels. The spatial filtered spectrum $\mathbf{I}^{spa}_{i,j}$ of the central pixel $\mathbf{I}^{spe}_{i,j}$ equals to the average of all spectra of green pixels. For a sample pixel $\mathbf{I}^{spe}_{i,j}$, we have the following equation in each band:

$$\mathbf{I}^{spa}_{i,j} = \frac{1}{h}\sum_{p,q} \mathbf{I}^{spe}_{p,q} \quad \text{if } \exists Y_{p,q} \text{ and } p \neq i, q \neq j \tag{1}$$

where h is the number of sample pixels in the window of $\mathbf{I}^{spe}_{i,j}$, and $Y_{p,q} \in \mathbb{R}$ is the label of pixel $\mathbf{I}^{spe}_{p,q}$. p and q are the coordinates relative to (i, j), and they are determined by window size. For example, when the window size is set to 3×3, p and q range from $i-1$ to $i+1$. We can temporarily assign a class to these samples when marking the samples in the target domain image because the target domain has no label information. For example, these samples are labeled class 1, whereas the nonsamples are labeled class 0 (indicating no label information). This class is only used to temporarily distinguish between sample and nonsample, and not for subsequent transfer learning classification.

The filter is applied to source and target domain images. After extracting the sample pixels, we obtain the original spectral features of source domain samples and target domain samples, $\mathbf{X}^{spe}_s \in \mathbb{R}^{n\times l}$ and $\mathbf{X}^{spe}_t \in \mathbb{R}^{m\times l}$, respectively. We can also obtain the spatial features of source domain samples and target domain samples, $\mathbf{X}^{spa}_s \in \mathbb{R}^{n\times l}$ and $\mathbf{X}^{spa}_t \in \mathbb{R}^{m\times l}$, respectively.

B. Manifold Embedded

To avoid feature distortions in the original feature space, we embed all samples into Grassmann manifold space by using the geodesic flow kernel (GFK) method to utilize the geometric structure of features in manifold space. This method assumes the PCA subspaces of the source domain and target domain as two different points in the manifold, and the geodesic flow between them can be regarded as transferring the source domain to target domain in manifold space. The subspace dimension is assumed to be d. Let $\mathbf{P}_s$ and $\mathbf{P}_t \in \mathbb{R}^{l\times d}$ denote the PCA subspaces of the source domain and target domain, respectively. $\mathbf{R}_s \in \mathbb{R}^{l\times(l-d)}$ denotes the orthogonal complement of $\mathbf{P}_s$. Then, the projection kernel $\mathbf{G} \in \mathbb{R}^{l\times l}$ can be obtained by the following:

$$\mathbf{G} = [\mathbf{P}_s\mathbf{U}_1\mathbf{R}_s\mathbf{U}_2]\begin{bmatrix}\mathbf{\Lambda}_1\mathbf{\Lambda}_2\\ \mathbf{\Lambda}_2\mathbf{\Lambda}_3\end{bmatrix}\begin{bmatrix}\mathbf{U}_1^T\mathbf{P}_s^T\\ \mathbf{U}_2^T\mathbf{R}_s^T\end{bmatrix} \tag{2}$$

where $\mathbf{U}_1 \in \mathbb{R}^{d\times d}$ and $\mathbf{U}_2 \in \mathbb{R}^{(l-d)\times d}$ are orthonormal matrices given by the singular value decomposition pair $\mathbf{P}_s^T\mathbf{P}_t = \mathbf{U}_1\mathbf{\Gamma}\mathbf{V}^T$ and $\mathbf{R}_s^T\mathbf{P}_t = -\mathbf{U}_2\mathbf{\Sigma}\mathbf{V}^T$. $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$, and $\mathbf{\Lambda}_3 \in \mathbb{R}^{d\times d}$ are diagonal matrices about the principal angles $\theta_i (i = 1, 2 \ldots, d)$ between $\mathbf{P}_s$ and $\mathbf{P}_t$. The diagonal elements are as follows:

$$\begin{aligned}\lambda_{1i} &= 1 + \frac{\sin(2\theta_i)}{2\theta_i}\\ \lambda_{2i} &= \frac{\cos(2\theta_i) - 1}{2\theta_i}\\ \lambda_{3i} &= 1 - \frac{\sin(2\theta_i)}{2\theta_i}.\end{aligned} \tag{3}$$

When spectral features are embedded into Grassmann manifold space, we use spectral features to construct kernel $\mathbf{G}^{spe}$. When spatial features are embedded into Grassmann manifold space, we use these spatial features to construct kernel $\mathbf{G}^{spa}$. Then, we use $\mathbf{Z} = \mathbf{X}\sqrt{\mathbf{G}^{spe}}$ to project $\mathbf{X}^{spe}_s$ and $\mathbf{X}^{spe}_t$ into manifold space and obtain the corresponding manifold spectral features of the source domain and target domain, $\mathbf{Z}^{spe}_s \in \mathbb{R}^{n\times l}$ and $\mathbf{Z}^{spe}_t \in \mathbb{R}^{m\times l}$, respectively. Similarly, the manifold spatial features of the source domain and target domain, $\mathbf{Z}^{spa}_s \in \mathbb{R}^{n\times l}$ and $\mathbf{Z}^{spa}_t \in \mathbb{R}^{m\times l}$, respectively, can be obtained after applying $\mathbf{Z} = \mathbf{X}\sqrt{\mathbf{G}^{spa}}$ to $\mathbf{X}^{spa}_s$ and $\mathbf{X}^{spa}_t$.

C. Kernel Construction

Kernel methods are widely used in transfer learning to solve the linearly inseparable problem in the original space, such as MMD. Many kernel methods only use a single kernel. However, some studies proved that a kernel constructed by combining several kernels may work better than a single kernel [48], [49]. The question is choosing and combining these kernels. Some methods consider integrating different kernel functions, such as radial basis function (RBF) kernel and linear kernel or uniting a kernel series with different parameters. In this article, we regard the manifold spatial features $\mathbf{Z}^{spa} = [\mathbf{Z}^{spa}_s, \mathbf{Z}^{spa}_t] \in \mathbb{R}^{(n+m)\times l}$ as the same data of manifold spectral features $\mathbf{Z}^{spe} = [\mathbf{Z}^{spe}_s, \mathbf{Z}^{spe}_t] \in \mathbb{R}^{(n+m)\times l}$ because they are assumed to be similar. However, they still exhibit some distinctions. Thus, their kernels can be treated as two different kernels of the same data, and the weighted kernel $\mathbf{K}^{ss} \in \mathbb{R}^{(n+m)\times(n+m)}$ can be written as follows:

$$\mathbf{K}^{ss} = \alpha\mathbf{K}^{spa} + (1-\alpha)\mathbf{K}^{spe} \tag{4}$$

where $\mathbf{K}^{spe} \in \mathbb{R}^{(n+m)\times(n+m)}$ is the kernel of $\mathbf{Z}^{spe}$. Let $K(\cdot,\cdot)$ denote the chosen kernel function; we have $\mathbf{K}^{spe}_{ij} =$

$\mathrm{K}(z_i^{spe}, z_j^{spe})$ for every two samples z_i^{spe} and z_j^{spe} of $\mathbf{Z}^{spe}$. Similarly, $\mathbf{K}^{spa} \in \mathbb{R}^{(n+m)\times(n+m)}$ is the kernel of $\mathbf{Z}^{spa}$ with $\mathbf{K}_{ij}^{spa} = \mathrm{K}(z_i^{spa}, z_j^{spa})$ for every two samples z_i^{spa} and z_j^{spa} of $\mathbf{Z}^{spa}$. α is the weights between two kernels because their importance is different, as defined by the average correlation coefficient of all corresponding sample pairs of $\mathbf{Z}_i^{spe}$ and $\mathbf{Z}_i^{spa}$. Thus, α can be estimated as follows:

$$\alpha = \frac{1}{n+m}\sum_{i=1}^{n+m}\frac{\mathrm{Cov}\left(\mathbf{Z}_i^{spe}, \mathbf{Z}_i^{spa}\right)}{\sigma_{\mathbf{Z}_i^{spe}}\sigma_{\mathbf{Z}_i^{spa}}} \tag{5}$$

where $\mathrm{Cov}(\cdot,\cdot)$ is the covariance of data pair, and σ is the standard deviation. A more effective kernel is constructed in this approach, fully utilizing spectral and spatial information.

D. Classifier Construction

Summarizing SRM, DDA, and the Laplacian matrix of manifold data, the classifier f can be represented as follows:

$$\begin{aligned} f = \underset{f\in\sum_{i=1}^{n} H_{\mathbf{K}}}{\arg\min} & \sum_{i=1}^{n}\left(y_i - f\left(z_i^{spe}\right)\right)^2 + \eta\|f\|_K^2 \\ & + \rho R_f\left(\mathbf{Z}_s^{spe}, \mathbf{Z}_t^{spe}\right) + \lambda\overline{D_f}\left(\mathbf{Z}_s^{spe}, \mathbf{Z}_t^{spe}\right) \end{aligned} \tag{6}$$

where $H_{\mathbf{K}}$ is the Hilbert space induced by the kernel $\mathbf{K}^{ss}$, y_i is the label of the ith manifold spectral features z_i^{spe}, and $\|f\|_{\mathbf{K}}^2$ is the squared norm of f. The first two terms are the application of SRM on the source domain. SRM can only be used on the source domain because no labeled information is available on the target domain. $R_f(\cdot,\cdot)$ is the regularization term that takes advantage of the geometric structure of the manifold. $\overline{D_f}(\cdot,\cdot)$ denotes the term of distribution alignment. λ, η, and ρ are the regularization coefficients of each term.

In the context of the representation theorem [50], f in (6) can be expressed as follows:

$$f(z) = \sum_{i=1}^{n+m}\boldsymbol{\beta}_i\mathbf{K}_{z_i,z}^{ss} \tag{7}$$

where $\boldsymbol{\beta}_i$ is the coefficient of the constructed kernel element $\mathbf{K}_{z_i,z}^{ss}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots)^T \in \mathbb{R}^{(n+m)\times C}$ be the coefficient matrix of $\mathbf{K}^{ss}$. Then, (6) can be rewritten as follows:

$$\begin{aligned} f = \underset{f\in\sum_{i=1}^{n} H_K}{\operatorname{argmin}} & \left\|\left(\mathbf{Y} - \boldsymbol{\beta}^T\mathbf{K}^{ss}\right)\mathbf{A}\right\|_F^2 + \eta\mathrm{tr}\left(\boldsymbol{\beta}^T\mathbf{K}^{ss}\boldsymbol{\beta}\right) \\ & + \rho\mathrm{tr}\left(\boldsymbol{\beta}^T\mathbf{K}^{ss}\mathbf{L}\mathbf{K}^{ss}\boldsymbol{\beta}\right) + \lambda\mathrm{tr}\left(\boldsymbol{\beta}^T\mathbf{K}^{ss}\mathbf{M}\mathbf{K}^{ss}\boldsymbol{\beta}\right). \end{aligned} \tag{8}$$

Each term in (8) has one-to-one correspondence to the terms in (6). The details are shown as follows.

The first two terms represent the SRM for minimizing the classification error, where $\mathbf{Y} = [y_1, \ldots, y_{n+m}] \in \mathbb{R}^{C\times(n+m)}$ is the label information corresponding to $\mathbf{Z}^{spe}$. We can only apply SRM to the source domain $\mathbf{Z}_s^{spe}$ because the label information of the target domain $\mathbf{Z}_t^{spe}$ is unknown. Thus, we set $\mathbf{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ as an indicator matrix whose elements are 0 except that the first n elements of the main diagonal are equal to 1. $\mathrm{tr}(\cdot)$ is the sum of the diagonal elements.

The third term is LR, which aims at exploring the role of manifold geometric structure in transfer learning. Hence, the Laplacian matrix $\mathbf{L}$ can be introduced as follows:

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \tag{9}$$

where $\mathbf{W}$ is the similarity matrix of data pairs. If $\mathbf{Z}_i^{spe}$ is the predetermined adjacent N points of $\mathbf{Z}_j^{spe}$ or the contrary, then $\mathbf{W}_{ij}$ is their cosine similarity; otherwise $\mathbf{W}_{ij}$ is 0. $\mathbf{D}$ is a diagonal matrix with elements $\mathbf{D}_{ij} = \sum_{j=1}^{n+m}\mathbf{W}_{ij}$.

The final term is the distribution alignment term, where M is the MMD matrix merged by the MMD matrices of the marginal distribution $\mathbf{M}_0$ and conditional distribution $\sum_{c=1}^{C}\mathrm{M}_c$, namely

$$\mathbf{M} = \mu\mathbf{M}_0 + (1-\mu)\sum_{c=1}^{C}\mathbf{M}_c. \tag{10}$$

The elements of $\mathbf{M}_0$ and $\mathbf{M}_c$ are as follows:

$$(\mathbf{M}_0)_{i,j} = \begin{cases} \frac{1}{n^2}, & z_i, z_j \in Z_s^{spe} \\ \frac{1}{m^2}, & z_i, z_j \in Z_t^{spe} \\ -\frac{1}{nm}, & \text{otherwise} \end{cases} \tag{11}$$

$$(\mathbf{M}_c)_{i,j} = \begin{cases} \frac{1}{n_c^2}, & z_i, z_j \in Z_s^{spe^{(c)}} \\ \frac{1}{m_c^2}, & z_i, z_j \in Z_t^{spe^{(c)}} \\ -\frac{1}{n_c m_c}, & \begin{cases} z_i \in Z_s^{spe^{(c)}}, & z_j \in Z_t^{spe^{(c)}} \\ z_i \in Z_t^{spe^{(c)}}, & z_j \in Z_s^{spe^{(c)}} \end{cases} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where n_c and m_c are the sample number of class c in the source and target domains. We train a simple classifier to obtain the soft labels of the target domain because the labels of the target domain are unknown. Thus, the conditional distribution can be estimated by the soft labels. Since the importance of marginal distribution and conditional distribution between the two domains is not always the same, it is necessary to weight them. On this foundation, we train a simple binary classifier h to distinguish source domain samples from target ones and introduce A distance, $d_A = 2(1-2\varepsilon(h))$, which is defined by the error of the classifier $\varepsilon(h)$. Then, the tradeoff parameter μ between $\mathbf{M}_0$ and $\mathbf{M}_c$ can be represented as follows:

$$\begin{aligned} \hat{\mu} &\approx \frac{\sum_{c=1}^{C} d_{\mathbf{M}_c}}{d_{\mathbf{M}_0} + \sum_{c=1}^{C} d_{\mathbf{M}_c}} \\ &= \frac{\sum_{c=1}^{C}\left(1 - 2\varepsilon\left(h_{\mathbf{M}_c}\left(\mathbf{Z}_s^{spe^{(c)}}, \mathbf{Z}_t^{spe^{(c)}}\right)\right)\right)}{1 - 2\varepsilon\left(h_{\mathbf{M}_0}\left(\mathbf{Z}_s^{spe}, \mathbf{Z}_t^{spe}\right)\right) + \sum_{c=1}^{C}\left(1 - 2\varepsilon\left(h_{\mathbf{M}_c}\left(\mathbf{Z}_s^{spe^{(c)}}, \mathbf{Z}_t^{spe^{(c)}}\right)\right)\right)}. \end{aligned} \tag{13}$$

By calculating the partial derivative of $\boldsymbol{\beta}$ for (13) and setting it to 0, we can obtain the solution of $\boldsymbol{\beta}$. Then, the final solution can be summarized as follows:

$$f = \mathbf{K}^{ss}\left(\left((\mathbf{A} + \lambda\mathbf{M} + \rho\mathbf{L})\mathbf{K}^{ss} + \eta\mathbf{I}\right)^{-1}\mathbf{A}\mathbf{Y}^T\right) \tag{14}$$

where $\mathbf{I} \in \mathbb{R}^{(n+m)\times(n+m)}$ is an identity matrix, and f is an $(n+m)\times C$ matrix. Each row represents a sample, and each column represents a class. Elements in the matrix represent the possibility of the corresponding sample belonging to the class. We consider the highest possibility class of each sample as its label and achieve the predicted labels of the target domain.

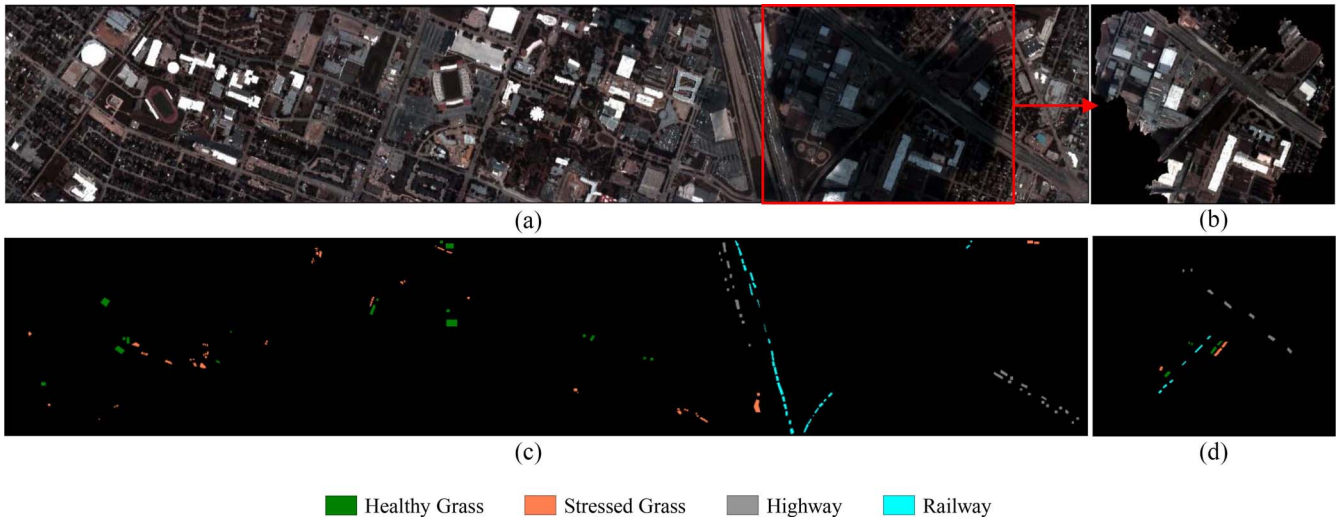

Fig. 3. (a) Image of the bright part of the Houston dataset (R:50, G:30, B:20). (b) Image of the shadow part of the Houston dataset (R:50, G:30, B:20). (c) Ground truth of the bright part. (d) Ground truth of the shadow part.

Algorithm 1 Procedure of the SSWK-MEDA Method

Input: Source domain image and its label $\mathbf{Y}_s$, target domain image, manifold subspace dimension d, regularization coefficient λ, η, ρ, repeat times T

Output: classifier f

Steps:

1. Apply the spatial filter (1) to the original source domain image and target domain image and extract samples to obtain spectral features $\mathbf{X}_s^{spe}$ and $\mathbf{X}_t^{spe}$, and spatial features $\mathbf{X}_s^{spa}$ and $\mathbf{X}_t^{spa}$.
2. Project spectral and spatial features into manifold space using (2), respectively, and obtain manifold spectral features $\mathbf{Z}^{spe}$ and manifold spatial features $\mathbf{Z}^{spa}$.
3. Construct the kernel $\mathbf{K}^{ss}$ by using (4) and calculate the adaptive factor α by using (5).
4. Train a simple classifier on $\mathbf{X}_s^{spe}$ to obtain the soft labels of $\mathbf{X}_t^{spe}$.
5. Calculate the Laplacian matrix $\mathbf{L}$ by using (9).
6. **Repeat**
7. Calculate the coefficient matrix of dynamic distribution alignment $\mathbf{M}$ by using (10).
8. Learn classifier f by using (14).
9. Update the soft labels of $\mathbf{X}_t^{spe}$.
10. **Until T times.**
11. Return classifier f.

IV. Experiments

A. Data Description

The first dataset is grss_dfc_2013 [24], [41], which is 349 × 1905 in size, with 144 bands covering the spectrum of 380–1050 nm. It was acquired by the National Center for Airborne Laser Mapping (NCALM), covering an area of the University of Houston with a spatial resolution of 2.5 m. The image is divided into two parts, namely, the bright and dark parts. The bright part has 4143 samples, and the dark part has 824 samples. The image and ground truth of the Houston dataset are shown in Fig. 3. Although the dataset has 12 classes, only four classes are shared by the two parts. Thus, we only use 4 of the 12 classes. These classes include healthy grass, stressed grass, highway, and railway. For convenience, we name this dataset as the Houston dataset.

The second dataset is the Indian Pines data set [51]–[53], which was produced by the airborne visible infrared imaging spectrometer (AVIRIS) sensor of the National Aeronautics and Space Administration. It includes 145 × 145 pixels with 200 bands covering the bandwidth of 375–2200 nm and a spatial resolution of 20 m. We divided it into two parts, namely, large and small parts. The large part has 4966 samples, and the small part has 1446 samples. The image and ground truth of the Indian Pines dataset are shown in Fig. 4. The experiment uses 7 out of the 16 classes, namely, corn-mintill, corn, soybeans-notill, soybeans-mintill, oats, bldg-grass-tree-drives, and hay-windrowed.

The third dataset is obtained by the Worldview-2 satellite, which has a spatial resolution of 1.8 m [41]. We called it as the Worldview dataset. It includes two images from two different periods of a region in Wuhan, namely, image 2011 and image 2012. Each image contains 200 × 200 pixels. It is a multispectral data, including eight bands (blue, green, yellow, red, coastal, near infrared, near infrared, and red edge). The four classes are as follows: 1) red building; 2) forest; 3) gray building; and 4) white building. There are 5317 samples in image 2011 and 5425 samples in image 2012. The image and ground truth of the Worldview dataset are shown in Fig. 5.

B. Experimental Setting

We construct six tasks for experiments denoted by "shadow→bright," "bright→shadow," "small→big," "big→small," "2011→2012," and "2012→2011." The source domain appears before the arrow "→" in each task, and the target domain appears after the arrow. The domains of each task belong to the same dataset. The experiments consider 1-nearest neighbor (1NN), SVM, SA, GFK, TCA, JDA, CORAL, and MEDA algorithms for comparison. Overall accuracy (OA) and kappa coefficient [22], [54], [55] are used as the evaluation indexes.

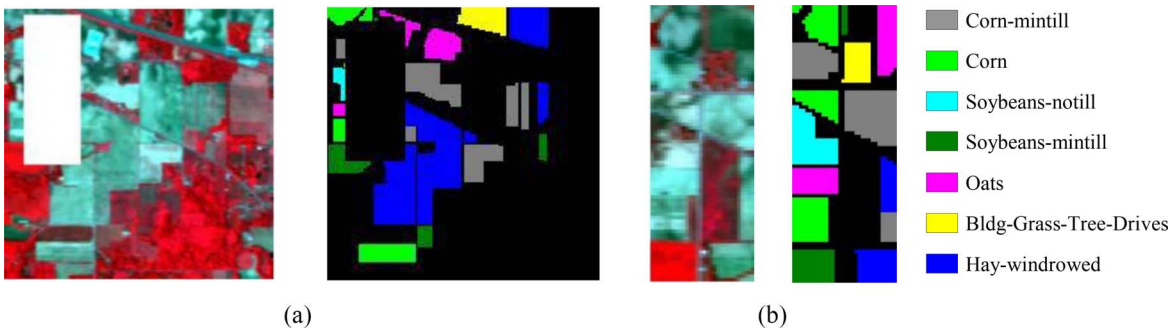

Fig. 4. Image (R:50, G:30, B:20) and ground truth of the Indian Pines dataset. (a) Big part. (b) Small part.

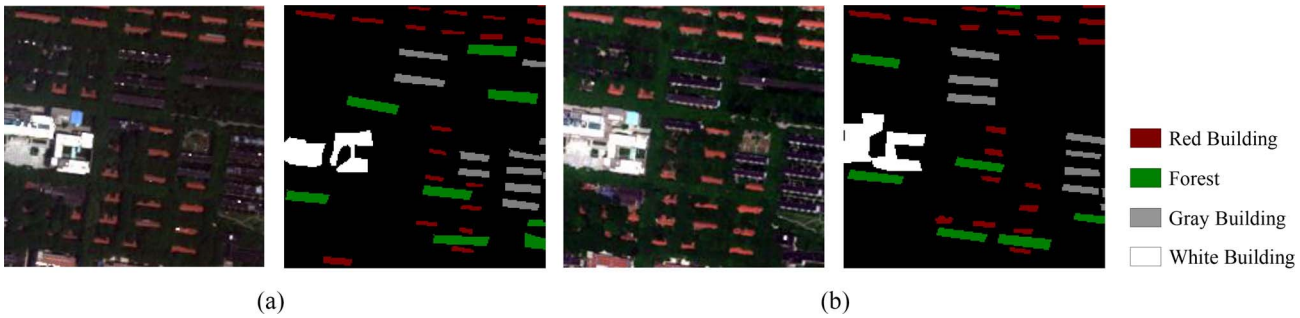

Fig. 5. Image (R:5, G:3, B:2) and ground truth of the subset of the Worldview dataset. (a) Time 2011. (b) Time 2012.

TABLE I
CLASSIFICATION ACCURACIES OF EACH COMPONENT OF SSWK-MEDA

Data set	Task	1NN		SRM		SRM+DDA		SRM+DDA+LR	
		OA(%)	kappa	OA(%)	kappa	OA(%)	kappa	OA(%)	kappa
Houston	bright→shadow	61.17	0.4595	78.64	0.6894	95.15	0.9326	**97.45**	**0.9646**
	shadow→bright	55.42	0.3983	76.88	0.6875	76.92	0.6882	**86.02**	**0.8146**
Indian Pines	big→small	25.31	0.1437	40.32	0.2853	58.99	0.5036	**59.54**	**0.5112**
	small→big	17.38	0.0908	35.12	0.2640	**51.53**	**0.4034**	**51.53**	**0.4035**
Worldview	2011→2012	63.82	0.5196	76.68	0.6842	86.91	0.8241	**91.45**	**0.8852**
	2012→2011	51.16	0.3731	46.79	0.3056	**97.29**	**0.9632**	**97.29**	**0.9632**

To obtain the soft labels of the target domain, we train a 1NN classifier based on source domain. We choose four nearest points as the adjacent points to calculate the matrix $\mathbf{W}$, which can achieve good effect. The chosen kernel function is RBF kernel, and its bandwidth is determined by the variance of inputs. Thus, the bandwidth of the RBF kernel of manifold spectral features is determined by the variance of manifold spectral features. The bandwidth of the RBF kernel of manifold spatial features is determined by the variance of manifold spatial features. Fig. 6(a) clearly shows that the accuracies of most tasks tend to converge when $T = 5$. Therefore, we set all repeat times T to 5.

1) Regularization Parameters: To evaluate the sensitivity of regularization parameters λ, η, and ρ, we ran SSWK-MEDA with their value from 0.0001 to 8. For convenience, the same regularization coefficients for each task are set to the same value in all iterations. The results in Fig. 6(b)–(d) show that the performance is not robust. When the values of η and ρ are small, their performance are robust. However, the performance of λ has no rules. Therefore, the parameters should be set in accordance with practical application.

Based on these results in Fig. 6, the values of these parameters are chosen. For the Houston dataset, the manifold subspace is set to $d = 60$, $\lambda = 1$, $\eta = 0.1$, and $\rho = 4$. For the Indian Pines dataset, the two tasks exhibit some differences. Thus, we set $d = 50, \lambda = 4$, $\eta = 0.0001$, and $\rho = 0.001$ for "big→small" and $\lambda = 0.1$, $\eta = 0.001$, and $\rho = 0.0001$ for "small→big." Similar to the Indian Pines dataset, the parameters of the two tasks of the Worldview dataset's subset are set differently, as follows: $d = 4$, $\lambda = 4$, $\eta = 0.001$, and $\rho = 0.1$ for "2011→2012" and $\lambda = 4$, $\eta = 0.01$, $\rho = 0.001$ for "2012→2011."

SSWK-MEDA includes three components, namely, SRM, DDA, and LR. Under the parameter setting, we conduct experiments to analyze the contributions of each component of SSWK-MEDA. Obviously, the results illustrated in Table I

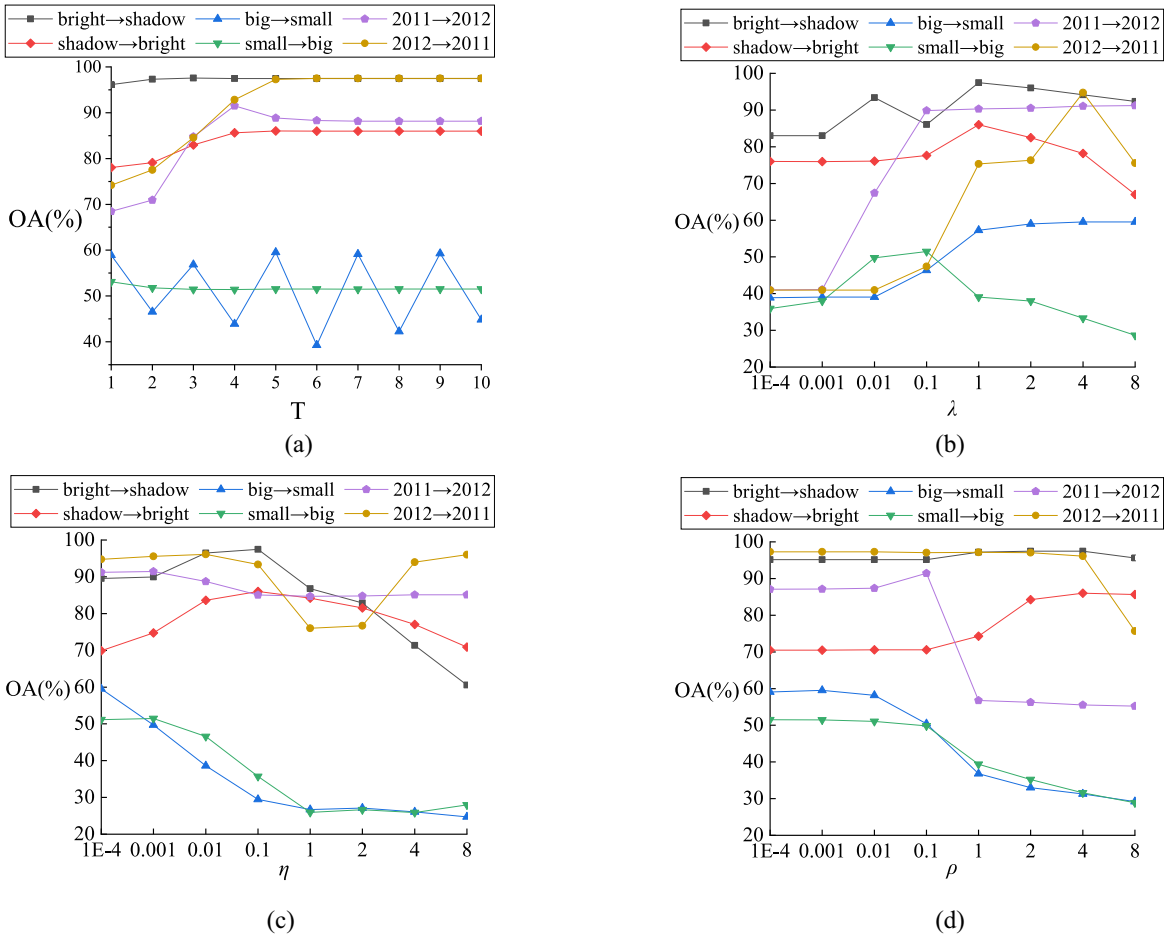

Fig. 6. OA of SSWK-MEDA under different values of (a) T, (b) λ, (c) η, and (d) ρ.

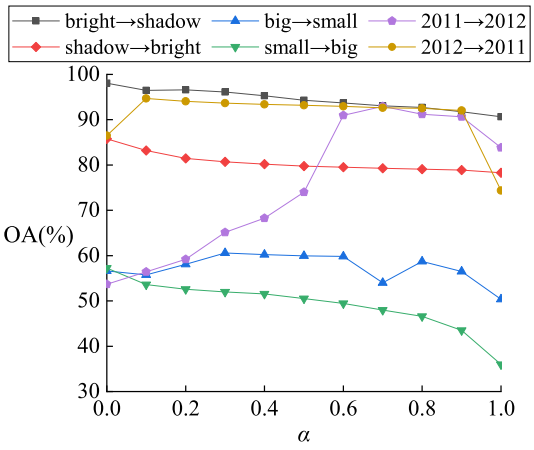

Fig. 7. OAs of SSWK-MEDA under different α values.

demonstrate that the components are effective in most tasks. These results also show that we need to set the parameters according to the actual situation.

2) Window Size: Different window sizes of spatial filter may cause different results. We conducted SSWK-MEDA under four different window sizes (3×3, 5×5, 7×7, and 9×9) to discuss the influence. The results in Table II clearly show that the OA trends of different window sizes do not have the same regularity among three datasets. All the results of different window sizes are still better than the other methods, which will be shown later. These results have verified the effectiveness of the spatial filter. Large window size may obtain better accuracy in some tasks, but may contain more dissimilar pixels in most cases, thereby reducing the accuracy. In addition, larger window size requires more time. Considering these factors comprehensively, we set the window size to 3×3 for all datasets.

3) Effectiveness of α: Evidently, different tasks have different optimal α values. Since the number of α is infinite, we take its value every 0.1 interval from 0 to 1 to observe the corresponding OA trends. Fig. 7 shows the corresponding OAs of different α values. We can clearly observe that only the tasks on the Houston dataset achieve robust performance. This finding has illustrated the need for adaptive weights. Adaptive weights can largely guarantee the accuracies of SSWK-MEDA toward different applications.

Different α values likely result in the same OA in some cases, such as "2011→2012." Thus, we only compare the corresponding OA. Let α_{opt} denote α, which obtain optimal OA under the previous setup, and α_{pre} denotes the predictive α of SSWK-MEDA. According to the results in Table III, the error between α_{opt} and α_{pre} is basically within $\pm 2.5\%$, except the task of small part→big part of the Indian Pines dataset. These results have testified the availability of SSWK-MEDA.

TABLE II
CLASSIFICATION ACCURACIES OF DIFFERENT SIZES OF WINDOWS ON DIFFERENT DATASETS

Data set	Task	3×3		5×5		7×7		9×9	
		OA(%)	kappa	OA(%)	kappa	OA(%)	kappa	OA(%)	kappa
Houston	bright→shadow	**97.45**	**0.9646**	94.05	0.9176	92.11	0.8906	91.02	0.8753
	shadow→bright	**86.02**	**0.8146**	85.06	0.8019	85.08	0.8022	85.37	0.8060
Indian Pines	big→small	59.54	0.5112	**64.38**	**0.5686**	62.59	0.5477	63.76	0.5643
	small→big	51.53	0.4035	**56.97**	**0.4504**	45.55	0.3229	39.21	0.2612
Worldview	2011→2012	91.45	0.8852	92.94	0.9055	96.15	0.9484	**97.99**	**0.9731**
	2012→2011	**97.29**	**0.9632**	92.08	0.8927	91.89	0.8901	94.58	0.9265

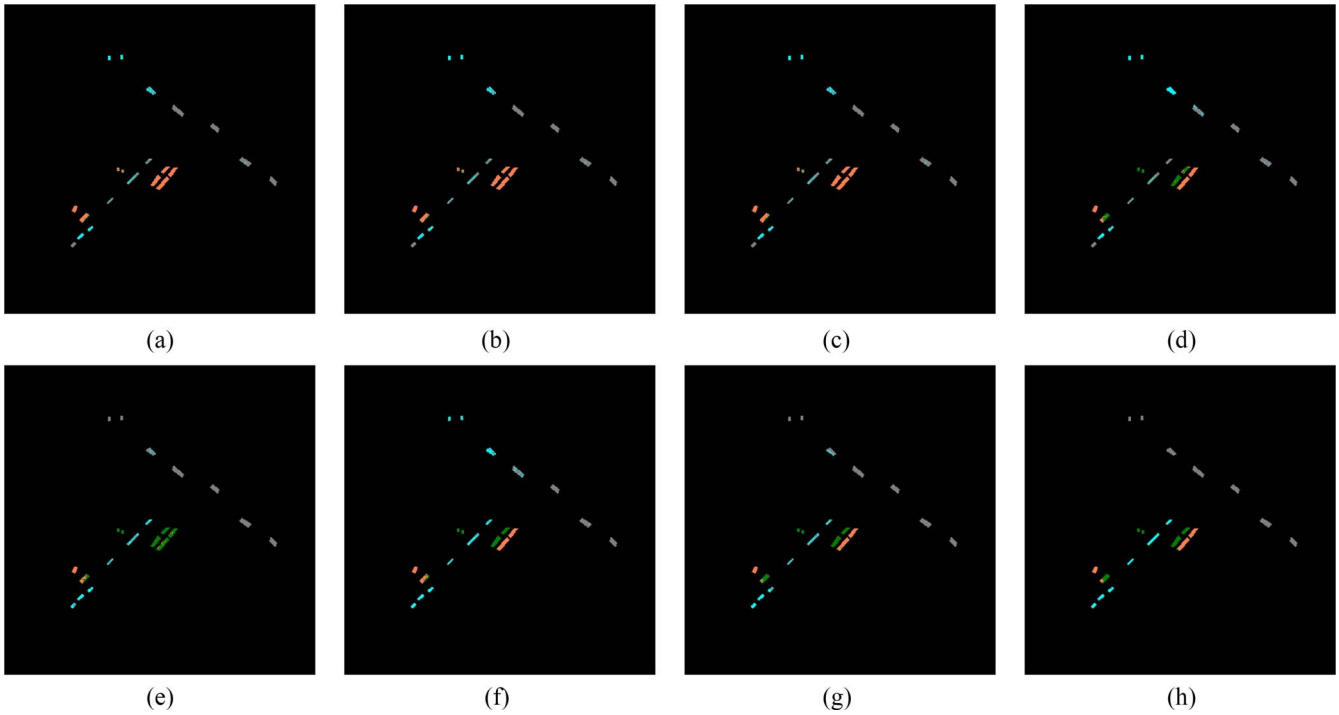

Fig. 8. Predicted labels of the shadow part of the Houston dataset (a) 1NN, (b) SA-1NN, (c) GFK-1NN, (d) CORAL-1NN, (e) TCA-1NN, (f) JDA-1NN, (g) MEDA, and (h) SSWK-MEDA.

TABLE III
ERROR BETWEEN PREDICTED α AND OPTIONAL α

Data set	Task	α_{opt}	α_{pre}	error
Houston	bright→shadow	98.06	97.45	- 0.61
	shadow→bright	85.78	86.02	+ 0.24
Indian Pines	big→small	60.58	59.54	- 1.04
	small→big	57.25	51.15	- 6.10
Worldview	2011→2012	92.96	91.12	- 1.84
	2012→2011	94.68	92.38	- 2.30

C. Performance

1) OA and Kappa: The OA and kappa classification accuracies with 1NN and SVM classifiers on the three datasets are illustrated in Tables IV–VII. We can obviously observe that in almost all tasks, SSWK-MEDA outperforms any other methods using 1NN or SVM classifier, including MEDA. And SSWK-MEDA improves at most 16.38% OA compared with the second-best methods.

The classification results of these tasks are demonstrated in Figs. 8–10. We can clearly see that the predicted images of SSWK-MEDA are more similar to the ground-truth image than the predicted image of other methods. In the Houston dataset, healthy grass and stressed grass are two very similar classes. Highway and railway are another similar classes in the Houston dataset. SSWK-MEDA can greatly distinguish them compared with other methods. For the Indian pines dataset, SSWK-MEDA achieves the highest accuracies of hay-windrowed. For the Worldview dataset, most methods inaccurately classify forest into gray building or cannot discriminate gray building from white building. SSWK-MEDA can handle this aspect elegantly. These results have demonstrated the effectiveness of SSWK-MEDA.

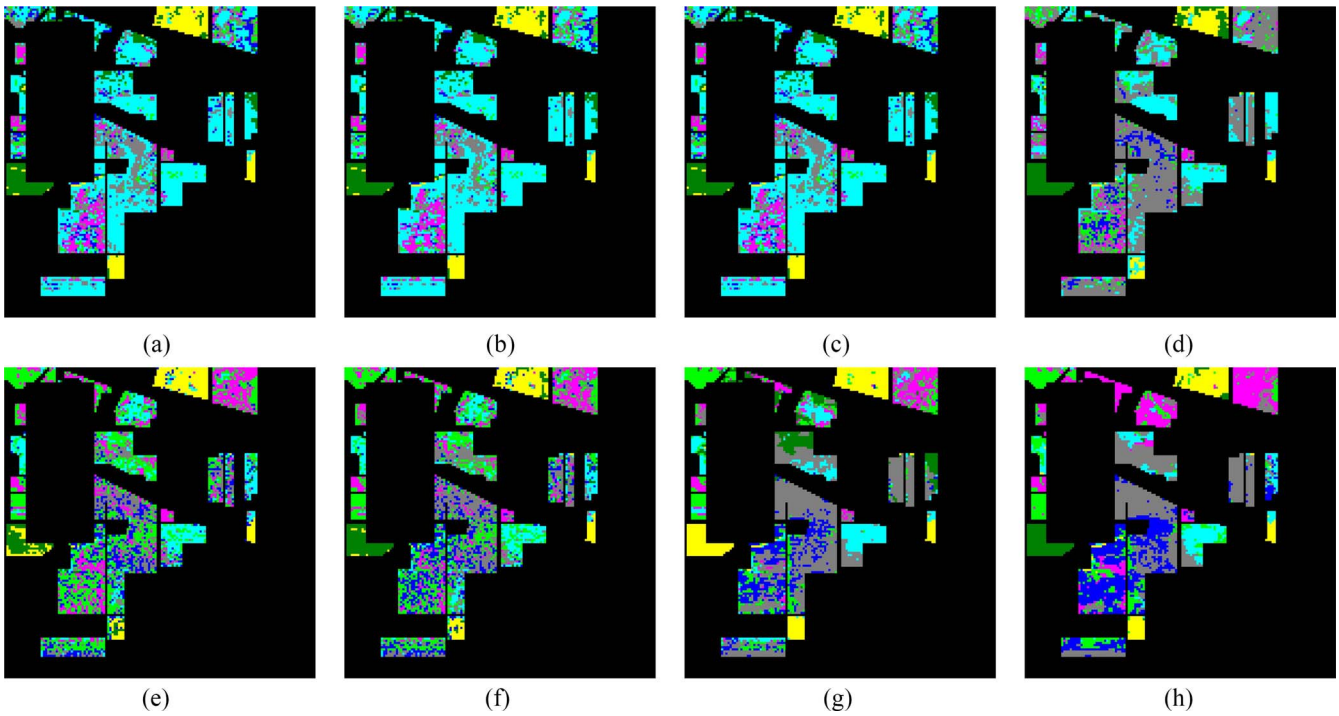

Fig. 9. Predicted labels of the big part of the Indian Pines dataset (a) 1NN, (b) SA-1NN, (c) GFK-1NN, (d) CORAL-1NN, (e) TCA-1NN, (f) JDA-1NN, (g) MEDA, and (h) SSWK-MEDA.

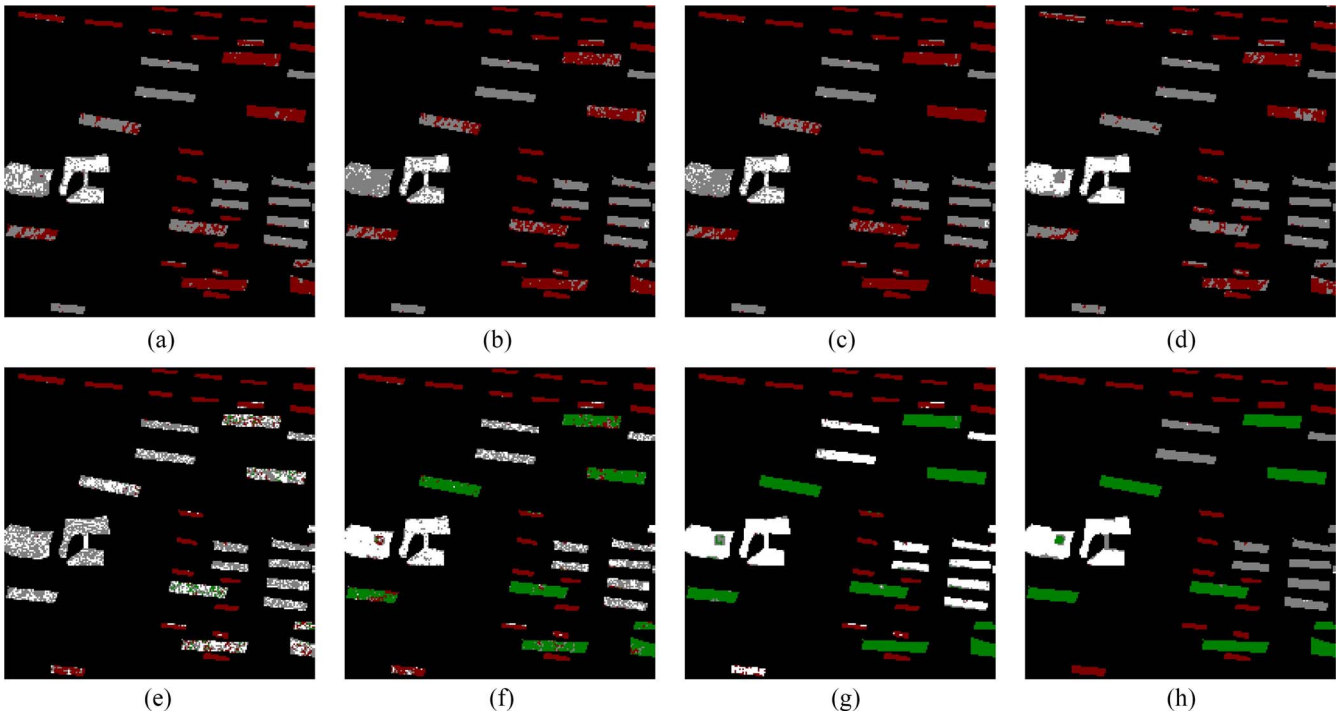

Fig. 10. Predicted labels of time 2011 of the Worldview dataset (a) 1NN, (b) SA-1NN, (c) GFK-1NN, (d) CORAL-1NN, (e) TCA-1NN, (f) JDA-1NN, (g) MEDA, and (h) SSWK-MEDA.

2) Time Complexity: We also empirically tested the time complexity of SSWK-MEDA and compared it with MEDA under the environment of Intel Core i7-8700 CPU and 16-GB RAM. The results in Table VIII show that SSWK-MEDA consumes slightly longer time than MEDA because of the few extra workloads. However, the difference in time consumption is unlikely to be large. Clearly, when larger window size of spatial filter is set, the SSWK-MEDA consumes longer time. However, compared with the overall time consumption, the additional time is immaterial. SSWK-MEDA consumes only a little more time than other methods, but the accuracy is

TABLE IV
OA(%) CLASSIFICATION ACCURACIES WITH 1NN CLASSIFIER ON DIFFERENT DATASETS

Data set	Task	1NN	SA-1NN	GFK-1NN	CORAL-1NN	TCA-1NN	JDA-1NN	MEDA	SSWK-MEDA
Houston	bright→shadow	61.17	61.29	61.29	72.09	76.46	82.40	91.02	**97.45**
	shadow→bright	55.42	55.39	55.42	70.48	62.01	63.43	78.30	**86.02**
Indian Pines	big→small	25.31	24.69	25.03	20.68	42.95	46.47	51.11	**59.54**
	small→big	17.38	15.67	16.35	25.71	30.65	35.04	35.92	**51.53**
Worldview	2011→2012	63.82	61.97	62.38	69.49	76.29	83.30	84.59	**91.45**
	2012→2011	51.16	48.07	48.99	55.33	51.93	80.91	74.27	**97.29**

TABLE V
OA(%) CLASSIFICATION ACCURACIES WITH SVM CLASSIFIER ON DIFFERENT DATA SETS

Data set	Task	SVM	SA-SVM	GFK-SVM	CORAL-SVM	TCA-SVM	JDA-SVM	MEDA	SSWK-MEDA
Houston	bright→shadow	64.44	63.96	63.96	72.69	71.12	81.43	91.14	**97.45**
	shadow→bright	57.23	34.25	57.42	65.82	61.33	85.69	79.39	**86.02**
Indian Pines	big→small	31.54	34.30	34.72	26.97	40.66	**63.07**	43.29	58.58
	small→big	16.15	16.83	16.79	19.29	35.22	36.79	32.66	**53.50**
Worldview	2011→2012	43.24	45.44	45.16	43.37	85.49	85.38	80.68	**91.65**
	2012→2011	46.21	49.75	47.77	51.70	54.79	72.58	74.97	**79.95**

TABLE VI
KAPPA CLASSIFICATION ACCURACIES WITH 1NN CLASSIFIER ON DIFFERENT DATASETS

Data set	Task	1NN	SA-1NN	GFK-1NN	CORAL-1NN	TCA-1NN	JDA-1NN	MEDA	SSWK-MEDA
Houston	bright→shadow	0.4595	0.4613	0.4613	0.6125	0.6688	0.7610	0.8742	**0.9646**
	shadow→bright	0.3983	0.3979	0.3983	0.6011	0.4861	0.5065	0.7102	**0.8146**
Indian Pines	big→small	0.1437	0.1373	0.1406	0.0902	0.3232	0.3588	0.4103	**0.5112**
	small→big	0.0908	0.0887	0.0884	0.1268	0.1889	0.2268	0.2238	**0.4035**
Worldview	2011→2012	0.5196	0.4613	0.4613	0.6125	0.6688	0.7610	0.8692	**0.8852**
	2012→2011	0.3731	0.3979	0.3983	0.6011	0.4861	0.5065	0.7102	**0.9632**

TABLE VII
KAPPA CLASSIFICATION ACCURACIES WITH SVM CLASSIFIER ON DIFFERENT DATASETS

Data set	Task	SVM	SA-SVM	GFK-SVM	CORAL-SVM	TCA-SVM	JDA-SVM	MEDA	SSWK-MEDA
Houston	bright→shadow	0.5096	0.5085	0.5083	0.6173	0.5818	0.7302	0.8759	**0.9646**
	shadow→bright	0.4222	0.1123	0.4248	0.5382	0.4771	0.8090	0.7233	**0.8146**
Indian Pines	big→small	0.2096	0.2363	0.2374	0.1475	0.3067	**0.5505**	0.3185	0.4984
	small→big	0.1033	0.0862	0.0853	0.0891	0.1955	0.2532	0.1816	**0.4134**
Worldview	2011→2012	0.2560	0.2852	0.2815	0.2577	0.8050	0.8037	0.7410	**0.8879**
	2012→2011	0.3122	0.3538	0.3315	0.3811	0.4079	0.6321	0.6623	**0.7266**

greatly improved, thereby illustrating that our method is worthy.

V. CONCLUSION

In this article, we propose a novel method called SSWK-MEDA for remote sensing image classification. SSWK-MEDA utilizes a filter to extract the mean spectrum of contiguous sample pixels of each sample pixel. The kernel of the background sample is regarded as the other kernel of the original sample, and they are adaptively weighed and integrated into a mixed kernel. Thus, the proposed method allows the use of spatial information to improve the classification accuracy. The proposed method can also eliminate the influence of feature

TABLE VIII
TIME CONSUMPTION (SECONDS) OF DIFFERENT METHODS

Data set	Task	1NN	SA-1NN	GFK-1NN	CORAL-1NN	TCA-1NN	JDA-1NN	MEDA	SSWK-MEDA
Houston	bright→shadow	**0.0622**	0.0823	0.1084	0.0812	12.9301	64.8780	16.4115	25.3474
	shadow→bright	**0.0596**	0.0774	0.1204	0.0767	13.1332	65.3389	16.4364	26.4909
Indian Pines	big→small	**0.1148**	0.1362	0.2135	0.1415	0.7202	240.7219	33.8293	35.0329
	small→big	**0.1250**	0.1416	0.2335	0.1473	0.7314	6.7986	33.8574	35.0398
Worldview	2011→2012	**0.2082**	0.2407	0.2264	0.2265	73.9478	375.6696	128.4368	130.3728
	2012→2011	**0.2078**	0.2275	0.2273	0.2273	74.1159	377.0714	128.2858	132.2123

distortion by using the geometric structure of manifold embedded data. We have run several experiments on hyperspectral data and multiple spectral data. All the results show that our method is more effective than the other existing methods, whether using a 1NN or SVM classifier. Future work will focus on leveraging the other relationship between pixels while minimizing the discrepancy of distributions.

ACKNOWLEDGMENT

The authors would like to thank Prof. L. Ma for generously providing them the Worldview-2 dataset for our experiment. The authors would also like to thank all the editors and anonymous reviewers for their careful reading and insightful remarks.

REFERENCES

[1] J. R. Anderson, *A Land Use and Land Cover Classification System for Use With Remote Sensor Data*, vol. 964. Washington, DC, USA: U.S. Govt. Publ., 1976.

[2] R. A. Schowengerdt, *Remote Sensing: Models and Methods for Image Processing*. San Diego, CA, USA: Academic, 2006.

[3] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[4] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sens. Environ.*, vol. 37, no. 1, pp. 35–46, Jul. 1991.

[5] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[6] 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.

[7] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, pp. 3–9, Dec. 2016.

[8] L. Bruzzone and M. Marconcini, "Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1108–1122, Apr. 2009.

[9] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[10] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together: A generalized framework for domain adaptation," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2144–2155, Jun. 2019.

[11] 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.

[12] Z. Deng, K.-S. Choi, Y. Jiang, and S. Wang, "Generalized hidden-mapping ridge regression, knowledge-leveraged inductive transfer learning for neural networks, fuzzy systems and kernel methods," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2585–2599, Dec. 2014.

[13] Z. Cui, W. Li, D. Xu, S. Shan, X. Chen, and X. Li, "Flowing on Riemannian manifold: Domain adaptation by shifting covariance," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2264–2273, Dec. 2014.

[14] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[15] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," *Knowl. Based Syst.*, vol. 80, pp. 14–23, May 2015.

[16] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in NLP," in *Proc. Ann. Meeting Assoc. Comput. Linguistics*, Prague, Czech Republic, 2007, pp. 264–271.

[17] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. Int. Conf. Mach. Learn.*, Corvallis, OR, USA, 2007, pp. 193–200.

[18] C. Persello, "Interactive domain adaptation for the classification of remote sensing images using active learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 736–740, Jul. 2013.

[19] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1410–1417.

[20] Y. Yang, C. Xu, R. Yang, and C. Meng, "Kernel extreme learning machine based domain adaptation," in *Proc. IEEE Int. Conf. Cloud Comput. Intell. Syst.*, Nanjing, China, 2018, pp. 593–597.

[21] 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.

[22] Q. Shi, B. Du, and L. Zhang, "Domain adaptation for remote sensing image classification: A low-rank reconstruction and instance weighting label propagation inspired algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5677–5689, Oct. 2015.

[23] X. Li, L. Zhang, B. Du, L. Zhang, and Q. Shi, "Iterative reweighting heterogeneous transfer learning framework for supervised remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 10, no. 5, pp. 2022–2035, May 2017.

[24] 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.

[25] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015.

[26] L. Bruzzone and D. F. Prieto, "Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 456–460, Feb. 2001.

[27] 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.

[28] J. Yang, R. Yan, and A. G. Hauptmann, "Adapting SVM classifiers to data with shifted distributions," in *Proc. IEEE Int. Conf. Data Min. Workshops*, Omaha, NE, USA, 2007, pp. 69–76.

[29] Y. Guo, X. Jia, and D. Paull, "A domain-transfer support vector machine for multi-temporal remote sensing imagery classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Fort Worth, TX, USA, 2017, pp. 2215–2218.

[30] Y. Guo, X. Jia, and D. Paull, "Effective sequential classifier training for SVM-based multitemporal remote sensing image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3036–3048, Jun. 2018.

[31] S. Xu, X. Mu, D. Chai, and S. Wang, "Adapting remote sensing to new domain with ELM parameter transfer," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1618–1622, Sep. 2017.

[32] S. Rajan, J. Ghosh, and M. M. Crawford, "Exploiting class hierarchies for knowledge transfer in hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3408–3417, Nov. 2006.

[33] M. Long, J. Wang, G. Ding, S. J. Pan, and S. Y. Philip, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1076–1089, May 2014.

[34] Z. Huang, Z. Pan, and B. Lei, "Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data," *Remote Sens.*, vol. 9, no. 9, p. 907, Sep. 2017.

[35] X. Ma *et al.*, "Regional atmospheric aerosol pollution detection based on LiDAR remote sensing," *Remote Sens.*, vol. 11, no. 20, p. 2339, Oct. 2019.

[36] 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.

[37] 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.*, 2013, pp. 2200–2207.

[38] 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 Int. Conf. Multimedia*, Seoul, South Korea, 2018, pp. 402–410.

[39] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 376–383.

[40] D. Tuia, J. Munoz-Mari, L. Gómez-Chova, and J. Malo, "Graph matching for adaptation in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 329–341, Jan. 2013.

[41] L. Ma, M. M. Crawford, L. Zhu, and Y. Liu, "Centroid and covariance alignment-based domain adaptation for unsupervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2305–2323, Apr. 2019.

[42] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput.*, 2013, pp. 2960–2967.

[43] H. Sun, S. Liu, S. Zhou, and H. Zou, "Unsupervised cross-view semantic transfer for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 13–17, Jan. 2016.

[44] H. Sun, S. Liu, S. Zhou, and H. Zou, "Transfer sparse subspace analysis for unsupervised cross-view scene model adaptation," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 9, no. 7, pp. 2901–2909, Jul. 2016.

[45] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2066–2073.

[46] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. Amer. Assoc. Artif. Intell. Conf.*, 2016, p. 8.

[47] 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.

[48] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[49] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[50] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[51] 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.

[52] Y. Dong, B. Du, L. Zhang, and L. Zhang, "Exploring locally adaptive dimensionality reduction for hyperspectral image classification: A maximum margin metric learning aspect," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 10, no. 3, pp. 1136–1150, Mar. 2017.

[53] Y. Zhou, J. Peng, and C. P. Chen, "Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 1082–1095, Feb. 2015.

[54] H. Luo, C. Liu, C. Wu, and X. Guo, "Urban change detection based on Dempster–Shafer theory for multitemporal very high-resolution imagery," *Remote Sens.*, vol. 10, no. 7, p. 980, Jul. 2018.

[55] 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.