Tensor Alignment Based Domain Adaptation for Hyperspectral Image Classification

Yao Qin, Student Member, IEEE, Lorenzo Bruzzone, Fellow, IEEE, and Biao Li

Abstract—This paper presents a tensor alignment (TA) based domain adaptation method for hyperspectral image (HSI) classification. To be specific, HSIs in both domains are first segmented into superpixels and tensors of both domains are constructed to include neighboring samples from single superpixel. Then we consider the subspace invariance between two domains as projection matrices and original tensors are projected as core tensors with lower dimensions into the invariant tensor subspace by applying Tucker decomposition. To preserve geometric information in original tensors, we employ a manifold regularization term for core tensors into the decomposition progress. The projection matrices and core tensors are solved in an alternating optimization manner and the convergence of TA algorithm is analyzed. In addition, a post-processing strategy is defined via pure samples extraction for each superpixel to further improve classification performance. Experimental results on four real HSIs demonstrate that the proposed method can achieve better performance compared with the state-of-the-art subspace learning methods when a limited amount of source labeled samples are available.

Index Terms—Domain adaptation (DA), hyperspectral image (HSI) classification, superpixel segmentation, tensor alignment

I. INTRODUCTION

IN THE past decades, extensive research efforts have been spent on hyperspectral remote sensing since hyperspectral data contains detailed spectral information measured in contiguous bands of the electromagnetic spectrum [1]–[3]. Due to the discriminative spectral information of such data, they have been used for a wide variety of applications, including agricultural monitoring [4], mineral exploration [5], and etc. One fundamental challenge in these applications is how to generate accurate land-cover maps. Although supervised learning for hyperspectral image (HSI) classification has been extensively developed in the literature (including random forest [6], support vector machine (SVM) [7], laplacian SVM (LapSVM) [8]–[10], decision trees [11] and support tensor machine (STM) [12]), sufficient labeled training samples should be available to obtain satisfactory classification results. This would require extensive and expensive field data collection campaigns. Furthermore, with the advance of newly-developed spaceborne hyperspectral sensors, large numbers of HSIs are easily collected and it is not feasible to timely label samples of the hyperspectral images as reference for training. Therefore, only limited labeled samples are available in most real applications of hyperspectral classification. According to the statistical theory in supervised learning, the data to be classified are expected to follow the same probability distribution function (PDF) of training data. However, since the physical conditions (i.e. illumination, atmosphere, sensor parameters, and etc.) can hardly be the same when collecting data, PDFs of training and testing data tend to be different (but related) [13]. Then how to apply the labeled samples of original HSI to the related HSI is challenging in such cases. These problems can be addressed by adapting models trained on a limited number of source samples (source domain) to new but related target samples (target domain). The problem should be further studied for the development of hyperspectral applications.

According to the machine learning and pattern recognition literature, the problem of adapting model trained on a source domain to a target domain is referred to as transfer learning or domain adaptation (DA) [13]. The main idea of transfer learning is to adapt the knowledge learned in one task to a related but different task. An excellent review of transfer learning can be found in [14], [15]. In general, transfer learning is divided into four categories based on the properties of domains and tasks, i.e. DA, multi-task learning, unsupervised transfer learning and self-taught learning. In fact, DA has greater impact on practical applications. When applied to classification problems, DA aims to generate accurate classification results of target samples by utilizing the knowledge learned on the labeled source samples. According to [3], DA techniques for remote sensing applications can be roughly categorized as selection of invariant features, adaptation of data distributions, adaptation of classifier and adaptation of classifier by active learning (AL).

In our case of HSI classification, we focus on the second category, i.e. adaptation of data distributions, in which data distributions of both domains are made as similar as possible to keep the classifier unchanged. Despite the fact that several DA methods have been proposed for HSI classification, they treat HSIs as several single samples, which renders them incapable of reflecting and preserving important spatial consistency of neighboring samples. In this paper, to exploit the spatial information in a natural and efficient way, tensorial processing is utilized, which treats HSIs as three-dimensional (3D) tensors. Tensor arithmetic is a generalization of matrix and vector arithmetic, and is particularly well suited to represent multilinear relationships that neither vector nor matrix algebra can capture naturally [16], [17]. The power of
tensortial processing for improving classification performance without DA has been proved in [18]–[22]. Similarly, when we apply tensorial processing to HSI in DA, multilinear relationships between neighboring samples in both HSIs are well captured and preserved, while conventional DA methods using vectorization deal with single samples. Tensor-based DA methods for visual application has demonstrated the efficacy and efficiency on the task of cross-domain visual recognition [17], whereas there are few published works on DA by using tensorial processing of HSIs.

To be specific, we propose a tensor alignment (TA) method for DA, which can be divided into two steps. First, the original HSI data cubes in both domains are divided into small superpixels and each central sample is represented as a 3D tensor consisting of samples in the same superpixel. In this way, each tensor is expected to include more samples from the same class. Since tensors are acted as “basic elements” in the TA method, we believe that high purity of tensors brings better adaptation performance. Second, taking into account the computational cost, we randomly select part of target tensors in the progress of TA to identify the invariance subspace between the two domains as projection matrices. This is done on the source and selected target tensors, and the subspace shared by both domains is obtained by utilizing three projection matrices \( \{ U^{(1)}, U^{(2)}, U^{(3)} \} \) with original geometry preserved. The solution is addressed through the Tucker decomposition [24] with orthogonal regularization on projection matrices and original tensors are represented by core tensors in the shared subspace. Fig. 1 illustrates the manifold regularized TA method with a 1-Nearest Neighbor (1NN) geometry preserved.

In addition to the TA method, after generating classification map, a post-processing strategy based on pure samples extraction of each superpixel is employed to improve performances. The pure samples in superpixels have similar spectral features and likely belong to the same class. Therefore, if most pure samples in a superpixel are classified as \( i \)-th class, it is probable that the remaining pure samples belong to the same class. Since samples in one superpixel may belong to two or even more classes and there are always classification errors in DA, the ratio of pure samples predicting as the same class might be reduced if we extract more pure samples. Therefore, we extract the pure samples by fixing the ratio as 0.7. Specifically, final pure samples are extracted by first projecting samples in each superpixel to principal component axis and then including more samples in the middle range of the axis till the ratio reaches 0.7. In this way the consistency of classification results on pure samples is enforced. To sum up, the main contributions of our research are presented in Section VI. Section VII summarizes the experimental datasets and setup. Results and discussions are presented in Section IV. Section V describes the contributions of our research.

II. RELATED WORK

This section briefly describes important studies related to the adaptation of data distributions, tensorial processing of hyperspectral data and basic concepts in multilinear algebra.

A. Adaptation of Data Distributions

Several methods for the adaptation of data distributions focus on subspace learning, where projected data from both domains are well aligned. Then, the same classifier (or regressor) is expected to be suitable for both domains. In [25], the data alignment is achieved through principal component analysis (PCA) or kernel PCA (KPCA). In [26], a PCA-based subspace alignment (SA) algorithm is proposed, where the source subspace is aligned as close as possible to the target subspace using a matrix transformation. In [17], features from convolutional neural network are treated as tensors and their invariant subspace is obtained through the Tucker decomposition. In [27], the authors align domains with canonical correlation analysis (CCA) and then perform change detection. The approach is extended to a kernel and semisupervised version in [28], where the authors perform change detection with different sensors. In [29], the supervised multi-view canonical correlation analysis ensemble is presented to address heterogeneous domain adaptation problems.
A few studies assume that data from both domains lie on the Grassmann manifold, where data alignment is conducted. In [30], the sampling geodesic flow (SGF) method is introduced and finite intermediate subspaces are sampled along the geodesic path connecting the source subspace and the target subspace. Geodesic flow kernel (GFK) method in [31] models infinite subspaces in the way of incremental changes between both domains. Along this line, GFK support vector machine in [32] shows the performance of GFK in nonlinear feature transfer tasks. A GFK-based hierarchical subspace learning strategy for DA is proposed in [33], and an iterative coclustering technique applied to the subspace obtained by GFK is proposed in [34].

Other studies hold the view that the subspace of both domains can be low-rank reconstructed or clustered. The reconstruction matrix is enforced to be low-rank and a sparse matrix is used to represent noise and outliers. In [35], a robust domain adaptation low-rank reconstruction (RDALRR) method is proposed, where a transformed intermediate representation of the samples in the source domain is linearly reconstructed by the target samples. In [36], the low-rank transfer subspace learning (LTSL) method is proposed where transformations are applied for both domains to resolve disadvantages of RDALRR. In [37], a low-rank and sparse representation (LRSR) method is presented by additionally enforcing the reconstruction matrix to be sparse. To obtain better results of reconstruction matrix, structured domain adaptation (SDA) in [38] utilizes block-diagonal matrix to guide iteratively the computation. Different from the above methods, latent sparse domain transfer (LSDT) in [39] is inspired by subspace clustering, while the low-rank reconstruction and instance weighting label propagation algorithm in [40] attempts to find new representations for the samples in different classes from the source domain by multiple linear transformations.

Other methods focus on feature extraction strategy by minimizing predefined distance measures, e.g., Maximum Mean Discrepancy (MMD) or Bregman divergence. In [41], transfer component analysis (TCA) tries to learn some transfer components across domains in a Reproducing Kernel Hilbert Space (RKHS) using MMD. It is then applied to remote sensing images in [42]. TCA is further improved by joint domain adaptation (JDA), where both the marginal distributions and conditional distributions are adapted in a dimensionality reduction procedure [43]. Furthermore, transfer joint matching (TJM) aims to reduce the domain difference by jointly matching the features and reweighting the instances across domains [44]. Recently, joint geometrical and statistical alignment (JGSA) is presented by reducing the MMD and forcing both projection matrices to be close [45]. In [46] and [47], the authors transfer category models trained on landscape views to aerial views for high-resolution remote sensing images by reducing MMD.

Different from the above category for feature extraction, several studies employ manifold learning to preserve the original geometry. In [48], both domains are matched through manifold alignment while preserving label (dis)similarities and the geometric structures of the single manifolds. The algorithm is extended to a kernelized version in [49]. Spatial information of HSI data is taken into account for manifold alignment in [50]. In [51], both local and global geometric characteristics of both domains are preserved and bridging pairs are extracted for alignment. In addition to manifold learning, the manifold regularized domain adaptation (MRDA) method integrates spatial information and the overall mean coincidence method to improve prediction accuracy [52]. Beyond classical subspace learning, manifold assumption and feature extraction methods, several other approaches are proposed in the literature, such as class centroid alignment [53], histogram matching [54] and graph matching [55].

B. Tensorial Processing of Hyperspectral Data

A few published works of tensor-based methods have been applied to HSI processing to fully exploit both spatial and spectral information. The texture features of HSI at different scales, frequencies and orientations are successfully extracted by the 3D discrete wavelet transform (3D-DWT) [56]. The gray level co-occurrence is extended to its 3D version in [57] to improve classification performance. Tensor discriminative locality alignment (TDLA) algorithm optimizes the discriminative local information for feature extraction [58], while local tensor discriminant analysis (LTDA) technique is employed in [59] for spectral-spatial feature extraction. The high-order structure of HSI along all dimensions is fully exploited by superpixel tensor sparse coding to better understand the data in [60]. Moreover, several conventional 2D methods are extended to the 3D for HSI processing, such as the 3D extension of empirical mode decomposition in [61], [62]. The modified tensor locality preserving projection (MTLPP) algorithm is presented for HSI dimensionality reduction and classification in [63].

C. Notations and basics of Multilinear Algebra

A tensor is a multi-dimensional array that generalizes matrix representation. Vectors and matrices are first and second order tensors, respectively. In this paper, we use lower case letters (e.g., \( x \)), boldface lowercase letters (e.g., \( x \)) and boldface capital letters (e.g., \( X \)) to denote scalars, vectors and matrices, respectively. Tensors of order 3 or higher will be denoted by boldface capital letters (e.g., \( X \)).

A \( k \)-th order tensor is denoted by \( \mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_M} \). The corresponding \( k \)-mode matricization of the tensor \( \mathcal{X} \), denoted by \( \mathbf{X}(k) \), unfolds the tensor \( \mathcal{X} \) with respect to mode \( k \). The operation between tensor \( \mathcal{X} \) and a matrix \( \mathbf{U}^{(k)} \in \mathbb{R}^{I_k \times I_k} \) is the \( k \)-mode product denoted by \( \mathcal{X} \times_{(k)} \mathbf{U}^{(k)} \), which is a tensor of size \( I_1 \times \cdots \times I_{k-1} \times I_k \times I_{k+1} \times \cdots \times I_M \). Briefly, the notion for the product of tensor with a set of projection matrices \( \{ \mathbf{U}^{(k)} \}_{k=1}^{M} \) excluding the \( l \)-mode is:

\[
\mathcal{X} \times_{(l)} \mathbf{U}^{(l)} = \mathcal{X} \prod_{k \neq l} \times_{(k)} \mathbf{U}^{(k)}
\]  

Tucker decomposition is one of the most well-known decomposition models for tensor analysis. It decomposes a \( M \) mode tensor \( \mathcal{X} \) into a core tensor \( \mathcal{G} \) multiplied by a set projection
matrices \( \{ U^{(k)} \}_{k=1}^{M} \) with the objective function defined as follows:

\[
\min_{G, \Omega} \| X - G \prod_{k=1}^{M} \times (k) U^{(k)} \|_F 
\]

(2)

where \( G \in \mathbb{R}^{J_1 \times \cdots \times J_M} \) and \( \Omega \) represents \( \{ U^{(k)} \}_{k=1}^{M} \) with \( U^{(k)} \in \mathbb{R}^{J_k \times J_k} \). For simplicity, we denote \( \prod_{k=1}^{M} \times (k) U^{(k)} \) as \( \| G; \Omega \| \).

By applying the \( l \)-mode unfolding, Eq. (2) can alternatively be written as

\[
\min_{G, \Omega} \| X_{(l)} - U^{(l)} G_{(l)} (U^{(-l)})^T \|_F 
\]

(3)

where \( G_{(l)} \) denotes the \( l \)-mode unfolding of \( G \), and \( U^{(-l)} \) denotes \( \prod_{k \neq l} \otimes U^{(k)} \). The vectorization of (3) can be formulated as

\[
\min_{G, \Omega} \| \text{vec}(X_{(l)}) - (U^{(-l)} \otimes U^{(l)}) \text{vec}(G_{(l)}) \|_F 
\]

(4)

Note that regularizations of \( G \) and \( \Omega \) are ignored in above equations.

III. PROPOSED TENSOR ALIGNMENT APPROACH

A. Problem Definition

Assume that we have \( n_s \) tensor samples \( \{ X_{ij}^{(s)} \}_{i=1}^{n_s} \) of mode \( M \) in the source domain, where \( X_{ij}^{(s)} \in \mathbb{R}^{J_1 \times \cdots \times J_M} \). Similarly, tensors in the target domain are denoted by \( \{ X_{ij}^{(t)} \}_{i=1}^{n_t} \). In this paper, we consider only homogeneous DA problem, thus we assume that \( X_{ij}^{(s)} \in \mathbb{R}^{J_1 \times \cdots \times J_M} \). In the context of DA, we follow the idea to represent the subspace invariance between two domains as projection matrices \( \Omega = \{ U^{(l)} \}_{l=1}^{M} \), where \( U^{(l)} \in \mathbb{R}^{J_l \times J_l} \). Intuitively, we propose to conduct subspace learning on the tensor samples in both domains with manifold regularization. Fig. 1 shows that the shared subspace is obtained by utilizing 3 projection matrices \( \{ U^{(1)}, U^{(2)}, U^{(3)} \} \). By performing Tucker decomposition simultaneously, the tensor samples in both domains are represented by the corresponding core tensors \( G_S = \{ G_{ij}^{(l)} \}_{l=1}^{n_s} \) and \( G_T = \{ G_{ij}^{(l)} \}_{l=1}^{n_t} \) with smaller dimensions in the shared subspace. The geometrical information should be preserved as much as possible via forcing manifold regularization during subspace learning. In the next subsection, we will introduce how to construct HSI tensors and perform tensor subspace alignment.

B. Tensors Construction

DA is achieved by tensor alignment, where HSI tensors are regarded as “basic elements”. Therefore, HSI tensors are expected to be as pure as possible. An ideal tensor should consist of samples belonging to the same class. However, spatial square patches centered at the training (or testing) samples can contain samples from different classes when extracted at the edge between different classes. To obtain pure tensors, superpixel segmentation [64] is performed on the first three principal components of the HSI. Fig. 2(a) shows the segmentation result of Pavia University data as an illustrative example. Then, samples surrounding the central sample in the same superpixel are utilized to form the tensor [see Fig. 2(b)]. This strategy can preserve the spectral-spatial structure of the HSI cube and takes into account the dissimilarity of various neighborhoods, particularly at the edges of different classes.

C. Method Formulation

After the construction of tensors, we weight matrices for both domains are computed to enforce manifold regularization. We compute the source weight matrix \( W_S \) in a supervised manner with labels of source tensors, while 10 nearest neighboring samples are searched via spectral angle measure (SAM) for target weight matrix \( W_T \). Then, the binary distance is employed for weight matrix construction. To reduce computation cost, dimensionality reduction for all tensors is conducted via multilinear PCA (MPCA). Given tensor samples and weight matrices for both domains, the optimal formulation problem is then defined as:

\[
\min_{G_S, G_T, \Omega} \sum_{i=1}^{n_s} \| X_{ij}^{(s)} - [G_{ij}^{(s)}; \Omega] \|_F^2 + \sum_{j=1}^{n_t} \| X_{ij}^{(t)} - [G_{ij}^{(t)}; \Omega] \|_F^2 + \lambda \sum_{i=1}^{n_s} \sum_{j=1}^{n_T} \| G_{ij}^{(s)} - G_{ij}^{(s)} \|_F^2 w_{ij}^{(s)} + \sum_{i=1}^{n_s} \sum_{j=1}^{n_T} \| G_{ij}^{(t)} - G_{ij}^{(t)} \|_F^2 w_{ij}^{(t)} \\
\text{s.t.} \quad U^{(l)^T} U^{(l)} = I, 1 \leq l \leq M
\]

(5)

where \( \lambda \) is a nonnegative parameter for controlling the importance of manifold regularization, \( w_{ij}^{(s)} \) and \( w_{ij}^{(t)} \) are weight between the \( i \)-th and \( j \)-th tensor in \( W_S \) and \( W_T \), respectively. When \( \lambda \) is small, the objective function depends mainly on the minimization of reconstruction errors for all tensor objects. When it is large, the objective function depends mainly on the preservation of tensor geometry information.

D. Optimization

The problem in (5) can be solved by alternatively updating \( \Omega \) and the cores \( G_S \) and \( G_T \) until the objective function converges. Herein, by applying the \( k \)-mode unfolding and according to (3), we obtain the following problem

\[
\min_{G_S, G_T, \Omega} \sum_{G \in \{ S, T \}} \sum_{i=1}^{n_G} \| X_{ij}^{(G)} - U^{(k)} G_{ij}^{(G)} (U^{(-k)})^T \|_F^2 + \lambda \sum_{G = \{ S, T \}} \sum_{i=1}^{n_G} \sum_{j=1}^{n_T} \| G_{ij}^{(G)} - G_{ij}^{(G)} \|_F^2 w_{ij}^{(G)}
\]

\[
+ \lambda \sum_{G = \{ S, T \}} \sum_{i=1}^{n_G} \sum_{j=1}^{n_T} \| G_{ij}^{(G)} - G_{ij}^{(G)} \|_F^2 w_{ij}^{(G)}
\]
s.t. \(U^{(l)T}U^{(l)} = I, 1 \leq l \leq M\) \hspace{1cm} (6)

where \(D\) is introduced to denote source (\(S\)) and target (\(T\)) domain for simplicity.

1) Updating \(G_S\) and \(G_T\): When \(U\) is fixed, by applying vectorization shown in (4), the problem is formulated as

\[
\begin{array}{l}
\min_{G_S, G_T} \sum_{G \in \{S, T\}} \sum_{i=1}^{n_G} \| \text{vec}(X_{G(k)}^i) - Z^{(k)} \text{vec}(G_{G(k)}^i) \|_F^2 \\
+ \lambda \sum_{G \in \{S, T\}} \sum_{i=1}^{n_G} \sum_{j=1}^{n_G} \| \text{vec}(G_{G(k)}^j) - \text{vec}(G_{G(k)}^i) \|_F^2 w_{ij}^G
\end{array}
\]  \hspace{1cm} (7)

where \(Z^{(k)} = U^{(-k)} \otimes U^{(k)}\). The matrix form of the above equation can be written as

\[
\begin{array}{l}
\min_{G_S, G_T} \sum_{G \in \{S, T\}} \| X_{G(k)}^U - Z^{(k)} G_{G(k)}^U \|_F^2 \\
+ \lambda \sum_{G \in \{S, T\}} \text{Tr}(G_{G(k)}^T L_{G}^D (G_{G(k)}^T)^T)
\end{array}
\]  \hspace{1cm} (8)

where \(X_{G(k)}^U\) and \(G_{G(k)}^U\) are matrices in which the \(i\)-th columns are \(\text{vec}(X_{G(k)}^i)\) and \(\text{vec}(G_{G(k)}^i)\), respectively. The \(L_{G}^D = D_{G}^D - W_{G}^D\) denotes the Laplacian matrix, \(D_{G}^D = \text{diag}(d_{G}^D, \ldots, d_{G}^D)\) and \(d_{G}^D = \sum_{j=1}^{n_G} W_{ij}^G\). Formally, we transform the problem above into the following optimization formulation

\[
\min_{G_S} \| X - ZG \|_F^2 + \lambda \text{Tr}(GLG^T)
\]  \hspace{1cm} (9)

For simplicity and better illustration, \(X \in \mathbb{R}^{n_x \times n_D}, Z \in \mathbb{R}^{n_x \times n_y}\) and \(L \in \mathbb{R}^{n_D \times n_p}\). Let \(\Lambda \Sigma \Sigma V^T\) be the singular value decomposition of \(Z\) and denote \(V^T G\) as \(Y\). Note that no information of \(G\) is lost in the transformation because \(V\) is invertible. Then we have

\[
\min_Y \| X - \Lambda \Sigma Y \|_F^2 + \lambda \text{Tr}(YY^T V^TV^T)
\]  \hspace{1cm} (10)

where \(Y \in \mathbb{R}^{n_y \times n_x}, A \Lambda^T = I_{n_y \times n_y}\) and \(V V^T = I_{n_y \times n_y}\). Based on the properties of trace and F-norm, we reformulate it as

\[
\min_Y \| M - \Sigma Y \|_F^2 + \lambda \text{Tr}(YY^T V^TV^T)
\]  \hspace{1cm} (11)

where \(M = \Lambda^T X\). We denote the \(i\)-th row of matrix \(M\) as \(M_i,:\). Then the problem above can be rewritten as

\[
\min_{Y_i} \sum_{i=1}^{n_y} \| M_i,: - \Sigma Y_i,: \|_F^2 + \lambda \sum_{i=1}^{n_y} Y_i,:) V Y_i,:^T
\]  \hspace{1cm} (12)

When only considering \(Y_i,:\), we have

\[
\min_{Y_i,:} \text{Q} Y_i,:^T - 2 \Sigma Y_i,:) M_i,: \hspace{1cm} (13)
\]

where \(Q = \lambda I + \Sigma^2 I_{n_D \times n_D}\) is a positive definite matrix. This is an unconstrained quadratic programming optimization of \(Y_i,:\) and can be easily solved by setting the derivation to zero. The optimal \(Y^*\) can be obtained by updating all rows and optimal \(G^*\) is given as \(VY^*\). When both \(G_{G(k)}^i\) are updated, the \(G_{G}^*\) can be obtained by applying tensorization.

2) Updating \(U\): When the cores \(G_S\) and \(G_T\) are fixed, we first write the problem of (5) as

\[
\min_{U(k)} \| X - G \prod_{l=1}^{M} U^{(l)T} U^{(l)} |^2_F, U^{(l)T} U^{(l)} = I, 1 \leq l \leq M
\]  \hspace{1cm} (14)

where \(X \in \mathbb{R}^{I_1 \times \cdots \times I_M \times (n_S + n_T)}\) and \(G \in \mathbb{R}^{I_1 \times \cdots \times I_M \times (n_S + n_T)}\) denote the concatenation of sample tensors \(X_{G}\) and core tensors \(G_{G}\) in each domain, respectively. Similar to most tensor decomposition algorithms, the solution for \(U^{(l)} (1 \leq l \leq M)\) is obtained by updating one with others fixed. By using the \(k\)-mode unfolding, the problem is derived as the following constrained objective function:

\[
\min_{U(k)} \| X^{(k)} - G U^{(k)} |^2_F, U^{(k)T} U^{(k)} = I
\]  \hspace{1cm} (15)

which can be effectively solved by utilizing singular value decomposition (SVD) of \(X^{(k)}\). Please refer to the Appendix for the proof. For an efficient computation, \(G(U^{(k)})^T\) is solved in the implementation as follows:

\[
G(U^{(k)})^T = [G \prod_{l \neq k} \times (l) U^{(l)}]_{\{k\}}
\]  \hspace{1cm} (16)

Based on the derived solutions of projection matrices and core tensors, the proposed method is summarized in Algorithm 1. Since the objective function in (5) is non-convex on projection matrices \(U\) and core tensors \(G_{G}\), we initialize projection matrices \(U\) to obtain a stationary solution by solving a conventional Tucker decomposition problem.

Algorithm 1 Tensor Alignment

**Input:** tensor set \(X_{G}\) in both domains, regularization parameter \(\lambda\) and dim of cores \(\{J_1, \ldots, J_M\}\)
**Output:** core tensor set \(G_{G}\) in both domains, projection matrices \(U\)

1. Compute two graph matrix \(W_{G}\);
2. Initialize \(U\) using Tucker decomposition;
3. While optimization in (5) does not converge do
   4. update \(G_{G}\) by solving (13);
   5. update \(U\) by alternatively solving (15);
   6. Check the convergence of (5);
7. Return \(G_{G}\) and \(U\).

E. Convergence and Computational Complexity Analysis

Formally, the objective function of the optimization in problem (5) is denoted as \(\Gamma(G_S, G_T, U)\). In (13), we update \(G_S\) and \(G_T\) with \(U\) fixed, i.e., we solve \(\{G_S^*, G_T^*\} = \text{argmin} \Gamma(G_S, G_T, U)\). Since both of them have a closed-form solution, we have \(\Gamma(G_S^*, G_T^*, U) \leq \Gamma(G_S, G_T, U)\) for any \(G_S, G_T, U\). Similarly, given the closed form solution of optimal \(U^*\), we have \(\Gamma(G_S^*, G_T^*, U^*) \leq \Gamma(G_S^*, G_T^*, U)\). Therefore, the \(\Gamma(G_S, G_T, U)\) decreases monotonically and iteratively, assuring the convergence of the proposed algorithm. As shown in Section VI, the proposed algorithm achieves convergence in less than 15 iterations.

The computational complexity mainly contains two parts: unconstrained optimization problem in (9) and orthogonal constrained problem in (15). The number of iterations for updating \(G_{G}\) is denoted as \(N_1\), while \(N_2\) is the average
number of iterations for updating $\Omega$ following each trial of updating $G_g$. For simplicity, the vectorization dimensionality of original and core tensors are denoted as $D_o = \prod_{k=1}^M I_k$ and $D_c = \prod_{k=1}^M J_k$, respectively. Firstly, the complexity of (9) consists of SVD of $Z$ in (10) and matrix inverse of $Q$ in (13). The corresponding complexities are $O(N_1D_oD_c^2)$ and $O[N_1D_o(n_S^2\log(n_S) + n_T^2\log(n_T))]$, respectively. Secondly, given the SVD of $X_{(k)}U^{(-k)}G^T$ ($k = 1, \ldots, M$) in (15), $\Omega$ is updated solved with complexity $O(N_1N_2(\sum_{k=1}^M I_kJ_k^2))$. In total, the complexity of $\text{TA}$ method is $O[N_1D_oD_c^2 + N_1D_o(n_S^2\log(n_S) + n_T^2\log(n_T)) + N_1N_2(\sum_{k=1}^M I_kJ_k^2)].$

IV. Pure Samples Based Classification Improvement

Once the projection matrices $\{U^{(1)}, U^{(2)}, U^{(3)}\}$ are computed, source and target tensors are represented as core tensors $G_S$ and $G_T$, respectively. The predicted map of target HSI can be easily obtained by a supervised classifier. It is notable that only part of target tensors are well exploited for domain adaptation and superpixel segmentation contributes only to tensor construction in the whole progress. In order to further exploit all target tensors and superpixel segmentation, in this section a strategy based on pure samples extraction is introduced to improve classification performance.

We firstly introduce the PCA-based method used for extracting the pure samples. As suggested in [65], for each superpixel we perform PCA and choose the first three principal components as the projection axis. Then, we project all samples onto these three principal components. For each projection axis, we normalize the projection values to pure $= \min p_i$. Further, $0$ is assigned to those pixels which meet $p_i \in (1 - T, T)$. In this way, each pixel is represented by three weights for three components. Finally, the sum of all weights for each sample is regarded as its purity index. The samples with purity index equal to $0$ are extracted as pure samples. Illustrative examples of pure samples extraction is shown in Fig. [3] where $T$ is set as $0.7$. After the extraction of pure samples in target HSI, we can apply the strategies for performance improvement.

The pure samples in each superpixel are expected to belong to the same class. However, there are always some samples predicting as different class in the testing stage. Therefore, if most of pure samples in one superpixel are predicted as $i$-th class, it is reasonable to believe the residual pure samples belong to the $i$-th class. Indeed, this idea is similar to the spatial filtering which also exploits spatial consistency. Since samples in one superpixel may belong to two even more classes and there are always classification errors in DA, the ratio of pure samples predicting as the same class might be reduced if we extract more pure samples. Therefore, we extract the pure samples by fixing the ratio of pure samples predicting as same class as $0.7$. If $70\%$ pure samples are predicted to belong to the $i$-th class, then remaining $30\%$ pure samples are changed into the $i$-th class. To find the optimal pure samples, a greedy algorithm is applied to extract more pure samples so that the ratio is no more than $0.7$:

$$N^*_\text{pure} = \min\{N_{\text{pure}} | \text{ratio}(N_{\text{pure}}) \leq 0.7\}$$

where $\text{ratio}(\cdot)$ means the ratio predicting as the same class. Intuitively, the optimal $N^*_\text{pure}$ should be as large as possible (to include more samples for the purpose of improving classification). Meanwhile, it should not be too large, otherwise samples belonging to different classes are included. We reduce the threshold $T$ by 0.01 iteratively to include more pure samples until the ratio of pure samples predicting as the same class reaches $70\%$. Then predicted results of remaining $30\%$ pure samples are alternated as the predicted class of the $70\%$ pure samples. We denote the strategy as $\text{TA}_P$ for short. Although the strategy is simple, experimental results in section V reveal that remarkable margins are gained by $\text{TA}_P$ over the proposed $\text{TA}$ method.

V. Data Description and Experimental Setup

A. DataSet Description

The first dataset consists in two hyperspectral images acquired by the Reflective Optics Spectrographic Image System

![Image](image-url)
(ROSIS) sensor over the University of Pavia and Pavia City Center are considered (see Fig. 4). The Pavia City Center image contains 102 spectral bands and has a size of 1096×492 pixels. The Pavia University image contains instead 103 spectral reflectance bands and has a size of 610×340 pixels. Seven classes shared by both images are considered in our experiments. The number of labeled pixels available is detailed in Table I. In the experiments, the Pavia University image was considered as the source domain, while the Pavia City Center image as the target domain, or vice versa. These two cases are denoted as univcenter and center/univ, respectively. Note that only 102 spectral bands of the Pavia University image were used for adaptation.

The second dataset is the GRSS2013 hyperspectral image consisting of 144 spectral bands. The data were acquired by the National Science Foundation (NSF)-funded Center for Airborne Laser Mapping (NCALM) over the University of Houston campus and the neighboring urban area. Originally, the data sets have a size of 1905×349 pixels and their 349 pixels, respectively. Note that spectral dimensionality setting as 20 λ.

The parameters of GFK, SA and TCA are tuned as in [31], [26] and [41], respectively. The dimension of final features in PCA is set as same as SA. The main parameters of the TA method are the window size, the tensor dimensionality after MPCA, the core tensor dimensionality after TA and the manifold regularization term λ. They are fixed as 5 × 5 pixels, 5 × 5 × 20, 1 × 1 × 10 and 1e-3 in all experiments, respectively. Note that spectral dimensionality setting as 20 and spatial dimensionality unchanged in MPCA guarantee that 99% energy is preserved and spatial information is also well kept, respectively.

### B. Experimental Setup

To investigate the classification performance of the proposed methods, SVM with linear kernel is employed as the supervised classifier. In detail, it is trained on the labeled source samples and tested on the unlabeled target samples. Although classifier like SVM with Gaussian kernel performs better in the classification task, the optimal parameters of such classifier tuned by source samples usually perform worse than expected for target samples under the context of DA. On the other hand, simple linear kernel is not biased by parameter tuning and can capture original relationships between samples from different domains. Free parameter C for linear SVM is tuned in the range (0.001-1000) by 5-fold cross validation.

Several unsupervised DA approaches for visual and remote sensing applications are employed as baseline methods:

- **Source (SRC):** SRC is the first baseline method that trains the classifier directly utilizing the labeled source samples.
- **Target (TGT):** TGT is the second baseline that trains the classifier directly utilizing the labeled target samples (upper bound on performance).
- **PCA:** PCA treats source and target samples as a single domain.
- **GFK:** GFK proposes a closed-form solution to bridge the subspaces of the two domains using a geodesic flow on a Grassmann manifold.
- **SA:** SA directly adopts a linear projection to match the differences between the source and target subspaces. Our approach is closely related to this method.
- **TCA:** TCA carries out adaptation by learning some transfer components across domains in a RKHS using MMD.

Given the computation cost of TA, we explore the adaptation ability of TA and TA_P with limited samples by randomly selecting tensors in both domains in each trial. To be specific, different numbers of tensors [5 10 15 20 25 30 35 40] for Pavia dataset and [3 4 5 6 7 8 9 10] for GRSS2013 dataset) per class from the source domain, and 100 tensors per class from the target domain are randomly selected for adaptation. After obtaining the projection matrices, SVM classifier with linear kernel is trained on selected source tensors and tested on all unlabeled target tensors. Regarding the SRC and TGT methods, central samples in selected source tensors and same number of samples per class randomly selected from target domain are employed for training, respectively. In the setting of other DA baseline methods, source tensors are vectorized as

| No. | Class     | Color in Fig. 5 | Pavia University | Pavia Center |
|-----|-----------|-----------------|------------------|--------------|
| 1   | Asphalt   | Yellow          | 547              | 704          |
| 2   | Meadows   | Green           | 569              | 685          |
| 3   | Trees     | Blue            | 451              | 793          |
| 4   | Soil      | Green           | 760              | 482          |
| 5   | Residential | Green       | 860              | 408          |
| 6   | Commercial| Purple          | 179              | 1065         |
| 7   | Road      | Brown           | 697              | 555          |
| 8   | Parking Lot 1 | Yellow     | 713              | 520          |

### Table I

**Number of Labeled Samples Available for the Pavia Dataset (top) and the GRSS2013 Dataset (down).**

- **TOP**:
- **SABELED**
- **ATASET**
- **D**
source samples and all target samples are used for adaptation. In the training stage, only the central samples are used as labeled. Take the case of 10 per class of source tensors as an example, then 250 (5 × 5 × 10) source samples per class are available for adaptation in DA datasets, and only 10 per class of central source samples are used for training the classifier. For each setting with same number of labeled source samples, 100 trials of the classification have been performed to ensure stability of the results. The classification results are evaluated using Overall Accuracy (OA), Kappa statistic and F-measure (harmonic mean of user’s and producer’s accuracies). All our experiments have been conducted by using Matlab R2017b in a desktop PC equipped with an Intel Core i5 CPU (at 3.1GHz) and 8GB of RAM.

VI. RESULTS AND DISCUSSIONS

A. Classification Performances

Tables II and III illustrate the OAs, Kappa statistics and the corresponding standard errors obtained by the proposed methods and the baseline methods for the Pavia and GRSS2013 datasets, respectively. In total, there are four cases: univcenter, centeruniv, left/right and right/left.

1) Pavia Dataset: Since more samples are used for adaptation and classifier training when increasing the number of labeled source samples from 5 to 40, the mean OAs and Kappa statistics of all methods roughly increase as expected. The increasing trend of mean OAs confirms that 100 trials are enough for achieving stable results. Moreover, standard errors of both OAs and Kappa statistics for small numbers of labeled samples appear to be higher. The mean OAs of TA and TA_P in univcenter case with various number of labeled samples are in the range of 74.3%-83.1% and 76.4%-86.1%, respectively. However, the performance in centeruniv case worsen for the proposed methods, with mean OA in the range of 52.8%-59.2% and 53.4%-58.7%, respectively. Similar trend can be found for other baseline methods. These results are not a surprise: the knowledge in Pavia University data can be easily transferred to Pavia Center data, whereas it’s not the same reversely. The mean OAs achieved by TGT for two cases are in the range of 82.1%-90.6% and 69.0%-81.9%, respectively. Roughly, the performance of TA_P in both cases is better than other baseline methods except TGT. More specifically, TGT yields 4.5%-5.7% and 7.5%-22.0% higher mean OAs than the TA_P for two cases, depending on the number of labeled samples. Compared with TA, mean OA achieved by TA_P are on average ~2.6% higher and ~0.1% lower for the univcenter and centeruniv case, respectively. Although the improvement for univcenter case is not so remarkable, this result confirms that introducing spatial consistency improves classification accuracy when the accuracy obtained by TA is not so small.

The accuracies achieved by SRC are about 4.6%-16.2%
These observations suggest that adaptation abilities of SRC i.e. 0%-3.2% and 0.1%-3.2% for center TCA accuracies of are globally minimized, poor performances are achieved by higher mean OAs (i.e. 5.8%-11.3% and 2.0%-3.0% for the two TA and 6.7%-19.2% lower than R Kappa). The accuracies achieved by SRC, the proposed Proposed TA_P methods for the univcenter case. However, when numbers of labeled samples per class is no less than 15, the differences between SRC and the proposed methods become lower for the centeruniv case, i.e. 0%-3.2% and 0.1%-3.2% for TA and TA_P, respectively. Further, SRC even outperforms TA and TA_P methods with numbers of labeled samples smaller than 15. One can see that the improvement of the proposed methods are relevant with number of labeled samples. The reason is that the adaptation ability of TA can be enhanced as expected with more samples. We can further notice that three methods (PCA, GFK and SA) perform better with more labeled samples for both cases. When comparing them with SRC, they roughly outperform SRC for two cases with enough labeled samples, whereas they perform worse with a small amount of labeled samples. These observations suggest that adaptation abilities of PCA, GFK and SA are severely hindered with a small number of source samples. The two proposed methods always deliver higher classification accuracies than all these three methods. As compared with TCA, the proposed TA_P method achieve higher mean OAs (i.e. 5.8%-11.3% and 2.0%-3.0% for the two cases). Since TCA seeks a new space where domain distances are globally minimized, poor performances are achieved by TCA with a small number of source samples, whereas the accuracies of TCA are better than other DA methods when increasing the number of labeled samples (see last column in Table II for the centeruniv case). It can be concluded that DA methods based on global alignment are affected by the availability of source samples.

2) GRSS2013 Dataset: Similar to the results of the Pavia dataset, the mean OAs and Kappa statistics of all methods increase as expected by increasing the number of labeled source samples, which further proves experimental stability. The mean OAs of TA and TA_P methods for left/right case with various numbers of labeled samples are in the range of 54.1%-60.4% and 54.4%-60.9%, respectively. However, both methods perform better in left/right case, with mean OAs in the range of 74.2%-83.0% and 74.6%-83.2%. By comparing classification performances of all methods, it is clear that GRSS2013 right data is easily transferred to left data. In fact, the difficulty of DA in left/right case lies in the shadow samples of the right dataset [see shadows in Fig. 5(b)]. The TA_P method in both cases outperforms most baseline DA methods, while TGT achieves the best accuracy for both cases Compared with TA, mean OAs achieved by TA_P are averagely ~0.4% and ~0.1% higher for the left/right and right/left cases, respectively.

The accuracies achieved by SRC are ~0.2% and ~0.4% smaller than those of TA and TA_P methods for the left/right case, respectively. Further, the differences become higher for the right/left case, i.e. 0.7%-3.3% and 1.1%-3.5% for TA and TA_P, respectively. Both TA and TA_P methods deliver higher classification accuracies than PCA, GFK and SA methods. It is further observed that TCA performs differently for the two
Fig. 6. Individual class accuracies (mean F-measure and standard errors) of different class (a-g) on the Pavia dataset. Top-row corresponds to the *univ*/center case, while down-row corresponds to the *center/univ* case (Best viewed in color).

Fig. 7. Individual class accuracies (mean F-measure and standard errors) of different class (a-h) on the GRSS2013 dataset. Top-row represents the *left/right* case, while down-row represents to the *right/left* case (Best viewed in color).

Fig. 8. Illustration of mean classification results for *univ*/center case by (a) *SRC* (Kappa = 0.605), (b) *TA* (Kappa = 0.797) and (c) *TA_P* (Kappa = 0.832).

B. Classification Maps and Individual Class Accuracies

To illustrate the effectiveness of the proposed *TA* and *TA_P* methods, Fig. 8 provides a comparison of mean classification maps over 100 trials obtained by the *SRC* and the proposed methods referring to the *univ*/center case with 40 labeled samples per class. In order to better display the classification results, average accuracy of each sample in the target image is computed and then transformed to the transparency. For example, when the average accuracy over 100 trials is 1, its transparency is set as 0. On the other hand, the transparency is set as 1 if the sample is never correctly classified. When blue color is applied as background in Fig. 8 deeper of blueness for one sample means being classified less accurately. Compared with *SRC* [Fig. 8(a)], local classification improvement of *TA* is easily observed [Fig. 8(b)]. The "asphalt" samples in the top-left (see the red rectangular box) and “bitumen” samples in the right-middle (see the square box) of Pavia center data are better classified by *TA* than *SRC*. Similarly, when comparing *TA* with *TA_P* [Fig. 8(c)], one can see that classification improvement of “asphalt” and “shadows” classes are achieved by *TA_P* [see red ellipses in Fig. 8(b-c)].

Fig. 9 reports individual class accuracies for the Pavia dataset, assessed by the mean F-measure (main curve) and its standard error (shaded area for each curve) over 100 trials. The results of 7 classes (“asphalt”, “meadows”, “trees”, “baresoil”, “bricks”, “bitumen”, “shadows”) are shown in Fig. 9(a-g), respectively. The top-row (a1-g1) corresponds to the *univ*/center case, while down-row (a2-g2) corresponds to the *center/univ* case. The following observations can be easily done:

- All methods perform more stably (smaller standard errors)
Fig. 9. Convergence curves of the proposed TA method in four cases with different number of labeled samples: (a) univ/center, (b) center/univ, (c) left/right, (d) right/left (Best viewed in color).

Fig. 10. Overall performance accuracy with different $d_2$ and $\lambda$ settings for four test cases: (a1-a3) univ/center; (b1-b3) center/univ; (c1-c3) left/right; (d1-d3) right/left. Each column corresponds to results on the same case with different window size (up to down: 3×3, 5×5, 7×7 pixels) (Best viewed in color).

with more labeled source samples in both cases. Roughly, most classes are classified more accurately with more labeled source samples except “bitumen” and “shadows” in univ/center case [see g(1) and f(1)]. The “bitumen” in both cases and “trees” in center/univ case have lowest accuracies than other classes, while “meadows” class in both cases has the best accuracy.

- The TGT method yields best results than other approaches for “asphalt”, “meadows”, “trees”, “bitumen” and “shadows” classes, whereas fails in detecting “bricks” class for univ/center case [see e(1)].
- The GFK method performs the worst for most classes in both cases, and has largest standard errors for “shadows” class in univ/center case. The PCA and SA methods behave similarly over all classes, whereas SRC performs worse for “bitumen” and “shadows” classes in univ/center case with more labeled samples.
- The TA and TA_P methods outperform other DA baselines on “trees”, “barensoil”, “bricks” and “bitumen” classes, and show comparable results on other classes, yielding a better overall classification accuracy. In addition, one can easily notice that a marginal yet obvious improvement by TA_P than TA on several classes, such as “bricks” and “bitumen”.

Referring to the GRSS2013 dataset, Fig. 7 illustrates individual class accuracies, where top-row [a(1)-h(1)] and down-row [a(2)-h(2)] represent left/right and right/left case, respectively. The results of 8 classes (“healthy grass”, “stressed grass”, “trees”, “soil”, “residential”, “commercial”, “road”, “parking lot”) are shown in Fig. 7(a-h), respectively. Similarly, conclusions related to the two proposed methods can be drawn. The accuracies of “healthy grass”, “trees” and “residential” achieved by TA and TA_P are higher than other DA baselines. TCA outperforms most methods on “commercial” class for left/right case. Considering that “commercial” is in the shadow area of right dataset, it is concluded that MMD measure is more suitable for large domain divergence.
C. Convergence

We have proved that the TA method is convergent under the iteratively updating rules of projection matrices and core tensors. Here, we investigate and demonstrate the speed of the convergence based on experimental results. Fig. 9 shows the mean reducing rate of the objective function of TA on the four test cases over 100 trials (shaded areas represent standard errors). One can see that the objective function value decreases by increasing the number of iterations. Moreover, we can observe that the TA converges very fast, usually taking less than 15 iterations. The reducing rate reaches 10e-6 and 10e-7 for the Pavia and the GRSS2013 datasets over all the 100 trials at the 15-th iteration, respectively. For our MATLAB implementation, 15 iterations take ~3.3s for univ/center cases when 40 labeled samples per class are used.

D. Parameter Sensitivity Analysis

Herein we perform experiments to discuss effects of parameters of the TA methods in the four cases. For simple and valid quantitative analysis, we only employ mean OAs to evaluate different parameter configurations. In detail, window size, spectral dimensionality of TA $d_2$ and manifold regularization $\lambda$ are discussed to achieve better understanding of the proposed method. Assuming that the window size is $W \times W$ pixels, we set tensor dimensions for MPCA and TA as $W \times W \times 20$ and $1 \times 1 \times d_2$, respectively. Note that the number of labeled samples are fixed to 40 per class for Pavia dataset and 10 per class for GRSS2013 dataset, and 10 trials are conducted for each sub-experiment. Fig. 10 illustrates the mean OAs with respect to different parameter configurations for the four cases. Each column indicates results using different window sizes for each case. One can observe that the trends of the mean OAs for all test cases under different window sizes are nearly the same, i.e. large value of $\lambda$ and low value of $d_2$ can both yield worse accuracy. The observation points out two conclusions: 1) large values of $\lambda$ force strong geometry preservation, hindering the learning of projection matrices; 2) If $d_2$ is smaller than 5 for all cases, no enough spectral information is preserved for training the classifier. However, when $d_2$ is larger than 10, there is no improvement of classification results. To sum up, $\lambda \in [0, 0.001]$ and $d_2 \in [10, 20]$ can be optimal parameter values for all cases.

VII. Conclusion

This paper has addressed the issue of DA in the classification of HSI under the assumption of small number of labeled samples. The main contributions of this paper are the proposed tensor alignment based domain adaptation algorithm and the strategy based on pure samples extraction for performance improvement.

The proposed TA naturally treats each sample in HSI as a 3-D tensor, exploiting the multilinear relationship between spatial and spectral dimensions. The shift between 3-D tensors from different domains is reduced by introducing a common set of projection matrices in the Tucker decomposition. The TA method mainly contains three steps, i.e. tensors construction, dimension reduction and tensor alignment. Firstly, HSIs in both domains are segmented into superpixels and each tensor is constructed by including samples from the same superpixel. In this way, the tensors are expected to contain samples belonging to the same class. Then, in order to reduce computational cost, MPCA is employed for spectral dimension reduction. In the stage of tensor alignment, to preserve the geometry of original tensors, two laplacian matrices from both domains are first computed. The problem of tensor alignment is formulated as jointly Tucker decomposition of tensors from both domains with manifold regularization on core tensors and orthogonal regularization on projection matrices. The solution is found by the developed efficient iterative algorithm and the convergence is analyzed. Once the projection matrices are computed, source and target tensors are represented as core tensors. The predicted map of target HSI can be easily obtained by a supervised classifier.

To further exploit the spatial consistency of HSI, a strategy for pure samples extraction for performance improvement is then proposed. The pure samples in each superpixel have similar spectral features and likely belong to the same class. Given that samples in one superpixel may belong to two or even more classes, pure samples may include samples belonging to different classes if we increase the number of pure samples extraction. To extract an appropriate number of pure samples, we fix the ratio of pure samples (predicted as the same class) as a constant value. We consider that it is reasonable to assume 70% pure samples in one superpixel. Although the strategy is simple, it turns out to be effective in performance improvement.

The experiments are conducted on four real HSIs, i.e. Pavia University and City Center, GRSS left and right images. To explore the adaptation capacity of TA, different numbers of source samples are randomly selected as labeled. Given the computational cost of TA, 100 tensors per class from target domain are selected for alignment. The TA method yields better results on univ/center, center/univ and right/left data sets than the other considered DA methods, whereas TCA outperforms TA on left/right data set. It is found that TCA performs better than all subspace learning method on the “Commercial” class in the left/right data set. Since the MMD-based TCA method can directly reduce the domain divergence, the “Commercial” class obtained under different conditions in the left/right data set (non-shadow and shadow area in the two images) is better adapted by TCA than by other considered DA methods. To summarize, the proposed TA method can achieve better performance compared with the state-of-the-art subspace learning methods when a limited amount of source labeled samples are available.

As future development, the proposed TA method can be easily extended to manifold regularization orthogonal Tucker decomposition for tensor data dimension reduction. Its typical applications include multichannel electroencephalographies, multiview images and videos processing.

ACKNOWLEDGMENT

The authors would like to thank Prof. P. Gamba from the University of Pavia for providing the ROSIS data, thank the
Hyperspectral Image Analysis Group and the NSF Funded Center for Airborne Laser Mapping (NCALM) at the University of Houston for providing the grss_dfe_2013 data set, and the IEEE GRSS Data Fusion Technical Committee for organizing the 2013 Data Fusion Contest.

APPENDIX A

PROOF OF SOLVING EQ. (15)

Theorem 1 Let $\text{ADV}^T$ be the singular value decomposition of $\text{AB}^T$, where $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times n}$. Then $X = \text{ADV} \times V^T$ is an orthogonal matrix minimizing $\|A - XB\|_F^2$, where $I_{m \times p}$ is a matrix with diagonal elements all 1 while others are 0.

Proof. To derive the method we first expand $\|A - XB\|_F^2$:

$$\|A - XB\|_F^2 = \|A\|_F^2 + \|B\|_F^2 - 2\text{tr}(X^T AB)^T$$

(18)

So picking $X$ to maximize $\text{tr}(X^T AB^T)$ will minimize $\|A - XB\|_F^2$. Let $\text{ADV}^T$ be the singular value decomposition of $\text{AB}^T$. Then we have:

$$\text{tr}(X^T AB^T) = \text{tr}(X^T \text{ADV}^T) = \text{tr}(V^T X^T \text{AD}^T)$$

(19)

Write $Z = V^T X^T A$, notice $Z$ is orthogonal (being the product of orthogonal matrices). The goal is re-stated: maximize $\text{tr}(ZD)$ through our choice of $X$. Since $D$ is diagonal, then $\text{tr}(ZD) = \sum_{i} z_{ii} d_{ii}$. The $D_{ii}$ are non-negative and $Z$ is orthogonal for any choice of $X$. The maximum is achieved by choosing $X$ such that all of $z_{ii} = 1$ which implies $Z = I_{m \times p}$. So an optimal $X$ is $\text{ADV} \times V^T$.

REFERENCES

[1] M. Faubel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, “Advances in spectral-spatial classification of hyperspectral images,” Proceedings of the IEEE, vol. 101, no. 3, pp. 652–675, 2013.

[2] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, “Advances in hyperspectral image classification: Earth monitoring with statistical learning methods,” IEEE signal processing magazine, vol. 31, no. 1, pp. 45–54, 2014.

[3] D. Tuia, C. Persello, and L. Bruzzone, “Domain adaptation for the classification of remote sensing data: An overview of recent advances,” IEEE geoscience and remote sensing magazine, vol. 4, no. 2, pp. 41–57, 2016.

[4] S. Schneider, R. J. Murphy, and A. Melkumyan, “Evaluating the performance of a new classifier—the gp-oad: A comparison with existing methods for classifying rock type and mineralogy from hyperspectral imagery,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 98, pp. 145–156, 2014.

[5] K. Tiwari, M. Arora, and D. Singh, “An assessment of independent component analysis for detection of military targets from hyperspectral images,” International Journal of Applied Earth Observation and Geoinformation, vol. 13, no. 5, pp. 730–740, 2011.

[6] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, “Investigation of the random forest framework for classification of hyperspectral data,” IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 3, pp. 492–501, 2005.

[7] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” IEEE Transactions on geoscience and remote sensing, vol. 42, no. 8, pp. 1778–1790, 2004.

[8] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,” Journal of machine learning research, vol. 7, no. Nov, pp. 2399–2434, 2006.

[9] S. Melacci and M. Belkin, “Laplacian support vector machines trained in the primal,” Journal of Machine Learning Research, vol. 12, no. Mar, pp. 1149–1184, 2011.

[10] W. Yang, X. Yin, and G-S. Xia, “Learning high-level features for satellite image classification with limited labeled samples,” IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 8, pp. 4472–4482, 2015.

[11] S. Delaibeux, B. Somers, B. Haest, T. Spanhove, J. V. Borre, and C. Mücher, “Heathland conservation status mapping through integration of hyperspectral mixture analysis and decision tree classifiers,” Remote sensing of environment, vol. 126, pp. 222–231, 2012.

[12] X. Guo, X. Huang, L. Zhang, L. Zhang, A. Plaza, and J. A. Benediktsson, “Support tensor machines for classification of hyperspectral remote sensing imagery,” IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 6, pp. 3248–3264, 2016.

[13] L. Bruzzone and M. Marconcini, “Domain adaptation problems: A dasvm classification technique and a circular validation strategy,” IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 5, pp. 770–787, 2010.

[14] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, “Visual domain adaptation: A survey of recent advances,” IEEE signal processing magazine, vol. 32, no. 3, pp. 53–69, 2015.

[15] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[16] M. A. O. Vasilescu and D. Terzopoulos, “Multilinear subspace analysis of image ensembles,” in Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 2. IEEE, 2003, pp. II–93.

[17] H. Lu, L. Zhang, Z. Cao, W. Wei, K. Xian, C. Shen, and A. van den Hengel, “When unsupervised domain adaptation meets tensor representations,” in The IEEE International Conference on Computer Vision (ICCV), vol. 2, 2017.

[18] Z. Zheng, B. Fan, J. Duan, L. Wang, K. Ding, S. Xiang, and C. Pan, “Discriminant tensor spectral–spatial feature extraction for hyperspectral image classification,” IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 5, pp. 1028–1032, 2015.

[19] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, “Compression of hyperspectral remote sensing images by tensor approach,” Neurocomputing, vol. 147, pp. 358–363, 2015.

[20] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, “Discriminative transfer subspace learning via low-rank and sparse representation,” IEEE Transactions on Image Processing, vol. 25, no. 2, pp. 850–863, 2016.

[21] Z. Feng, M. Wang, S. Yang, Z. Liu, L. Liu, B. Wu, and H. Li, “Superpixel tensor sparse coding for structural hyperspectral image classification,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 4, pp. 1632–1639, 2017.

[22] H. Lu, L. Zhang, Z. Cao, W. Wei, K. Xian, C. Shen, and A. van den Hengel, “When unsupervised domain adaptation meets tensor representations,” in The IEEE International Conference on Computer Vision (ICCV), vol. 2, 2017.

[23] Z. Zheng, B. Fan, J. Duan, L. Wang, K. Ding, S. Xiang, and C. Pan, “Discriminant tensor spectral–spatial feature extraction for hyperspectral image classification,” IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 5, pp. 1028–1032, 2015.

[24] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, “Compression of hyperspectral remote sensing images by tensor approach,” Neurocomputing, vol. 147, pp. 358–363, 2015.

[25] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, “Discriminative transfer subspace learning via low-rank and sparse representation,” IEEE Transactions on Image Processing, vol. 25, no. 2, pp. 850–863, 2016.

[26] Z. Feng, M. Wang, S. Yang, Z. Liu, L. Liu, B. Wu, and H. Li, “Superpixel tensor sparse coding for structural hyperspectral image classification,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 4, pp. 1632–1639, 2017.
[32] A. Samat, P. Gamba, J. Abuduwali, S. Liu, and Z. Miao, “Geodesic flow kernel support vector machine for hyperspectral image classification by unsupervised subspace feature transfer,” *Remote Sensing*, vol. 8, no. 3, p. 234, 2016.

[33] B. Banerjee and S. Chaudhuri, “Hierarchical subspace learning based unsupervised domain adaptation for cross-domain classification of remote sensing images,” *IEEE Journal of Selected Topics in Applied Intelligence and Remote Sensing and Remote Letters*, vol. 10, no. 11, pp. 5099–5109, 2017.

[34] B. Banerjee and K. M. Buddhiraju, “Domain adaptation in the absence of source domain labeled samples co-clustering-based approach,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1905–1909, 2016.

[35] I.-H. Juo, D. Liu, D. Lee, and S.-F. Chang, “Robust visual domain adaptation with low-reconstruction,” in *Computer Vision and Pattern Recognition (CVPR)*, 2012 IEEE Conference on. IEEE, 2012, pp. 2168–2175.

[36] M. Shao, D. Kit, and Y. Fu, “Generalized transfer subspace learning through low-rank constraint,” *International Journal of Computer Vision*, vol. 109, no. 1-2, pp. 74–93, 2014.

[37] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, “Discriminative transfer subspace learning via low-rank and sparse representation,” *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 850–863, 2016.

[38] J. Li, Y. Wu, and K. Lu, “Structured domain adaptation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1700–1713, 2017.

[39] L. Zhang, W. Zuo, and D. Zhang, “Lsdt: Latent sparse domain transfer learning,” *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1177–1191, 2016.

[40] Q. Shi, B. Du, and L. Zhang, “Domain adaptation for remote sensing image classification: A low-rank reconstruction and instance weighing label propagation inspired algorithm,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 10, pp. 5677–5689, 2015.

[41] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 190–210, 2011.

[42] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, “Semisupervised transfer component analysis for domain adaptation in remote sensing image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3550–3564, 2015.

[43] M. Long, J. Wang, G. Ding, J. Sun, and S. Y. Philip, “Transfer feature learning with joint distribution adaptation,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2200–2207.

[44] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer joint matching for unsupervised domain adaptation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1410–1417.

[45] J. Zhang, W. Li, and P. Oggunbona, “Joint geometrical and statistical alignment for visual domain adaptation,” arXiv preprint arXiv:1705.05498, 2017.

[46] H. Sun, S. Liu, S. Zhou, and H. Zou, “Transfer sparse subspace analysis for unsupervised cross-view scene model adaptation,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 7, pp. 2901–2909, 2016.

[47] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, “Semisupervised manifold alignment of multimodal remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 12, pp. 7708–7720, 2014.

[48] D. Tuia and G. Camps-Valls, “Kernel manifold alignment for domain adaptation,” *PloS one*, vol. 11, no. 2, p. e0148655, 2016.

[49] H. L. Yang and M. M. Crawford, “Spectral and spatial proximity-based manifold alignment for multitemporal hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 51–64, 2016.

[50] ——, “Domain adaptation with preservation of manifold geometry for hyperspectral image classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 543–555, 2016.

[51] C. Luo and L. Ma, “Manifold regularized distribution adaptation for cross-domain adaptation of remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, p. 329–341, 2013.

[52] L. Zhu and L. Ma, “Class centroid alignment based domain adaptation for classification of remote sensing images,” *Pattern Recognition Letters*, vol. 83, pp. 124–132, 2016.

[53] S. Inamdar, F. Bovolo, L. Bruzzone, and S. Chaudhuri, “Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 4, pp. 1243–1252, 2008.

[54] D. Tuia, J. Munoz-Mari, L. Gomez-Chova, and J. Malo, “Graph matching for adaptation in remote sensing,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 329–341, 2013.

[55] Y. Qian, M. Ye, and J. Zhou, “Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 2276–2291, 2013.

[56] F. Tsai and J.-S. Lai, “Feature extraction of hyperspectral image cubes using three-dimensional gray-level cooccurrence,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 6, pp. 3504–3513, 2013.

[57] L. Zhang, L. Zhang, D. Tao, and X. Huang, “Tensor discriminative locality alignment for hyperspectral image spectral–spatial feature extraction,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 242–256, 2013.

[58] Z. Zhong, B. Fan, J. Duan, L. Wang, K. Ding, S. Xiang, and C. Pan, “Discriminant tensor spectral–spatial feature extraction for hyperspectral image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1028–1032, 2015.

[59] Z. Feng, M. Wang, S. Yang, Z. Liu, L. Liu, B. Wu, and H. Li, “Superpixel tensor sparse coding for structural hyperspectral image classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 4, pp. 1632–1639, 2017.

[60] H. L. Yang and M. M. Crawford, “Spectral and spatial proximity-based manifold alignment for cross-domain adaptation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 11, pp. 6625–6643, 2016.

[61] Y.-J. Deng, H.-C. Li, L. Pan, L.-Y. Shao, Q. Du, and W. J. Emery, “Modified tensor locality preserving projection for dimensionality reduction of hyperspectral images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 8, 2016.

[62] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Sässstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[63] G. Martin and A. Plaza, “Spatial-spectral preprocessing prior to end-member identification and unmixing of remotely sensed hyperspectral data,” *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 5, no. 2, pp. 380–395, 2012.

Yao Qin (SM’ 16) received the B.S degree in Information Engineering from Shanghai Jiaotong University, Shanghai, China, in 2013 and the M. S. degree from School of Electronic and Engineering, National University of Defence Technology, Changsha, China, in 2015. Currently, he is pursuing the Ph. D. degree and is a visiting Ph. D at RSLab in the Department of Information Engineering and Computer Science, University of Trento, Italy. His major research interests are remote sensing image classification and domain adaptation.
Lorenzo Bruzzone (S’95–M’98–SM’03–F’10) received the Laurea (M.S.) degree in electronic engineering (summa cum laude) and the Ph.D. degree in telecommunications from the University of Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications at the University of Trento, Italy, where he teaches remote sensing, radar, and digital communications. Dr. Bruzzone is the founder and the director of the Remote Sensing Laboratory in the Department of Information Engineering and Computer Science, University of Trento. His current research interests are in the areas of remote sensing, radar and SAR, signal processing, machine learning and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the Principal Investigator of many research projects. Among the others, he is the Principal Investigator of the Radar for icy Moon exploration (RIME) instrument in the framework of the JUpiter ICy moons Explorer (JUICE) mission of the European Space Agency. He is the author (or coauthor) of 215 scientific publications in referred international journals (154 in IEEE journals), more than 290 papers in conference proceedings, and 21 book chapters. He is editor/co-editor of 18 books/conference proceedings and 1 scientific book. His papers are highly cited, as proven form the total number of citations (more than 25000) and the value of the h-index (74) (source: Google Scholar). He was invited as keynote speaker in more than 30 international conferences and workshops. Since 2009 he is a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS).

Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, July 1998. Since that he was recipient of many international and national honors and awards, including the recent IEEE GRSS 2015 Outstanding Service Award and the 2017 IEEE IGARSS Symposium Prize Paper Award. Dr. Bruzzone was a Guest Co-Editor of many Special Issues of international journals. He is the co-founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003 he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the founder of the IEEE Geoscience and Remote Sensing Magazine for which he has been Editor-in-Chief between 2013–2017. Currently he is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing. He has been Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society between 2012–2016.

Biao Li was born in Zhejiang, China, in 1968. He received the Ph.D. degrees in the School of Electronic Science, National University of Defense Technology, Changsha, China, in 1998, where he is currently a full professor. His current research interests include signal processing and infrared image processing.