A Multi-Kernel domain adaptation method for unsupervised transfer learning on cross-source and cross-region remote sensing data classification

Wei Liu and Rongjun Qin, *Senior Member, IEEE*

Abstract—Labeling remote sensing data for classification is labor-intensive and time-consuming in practical applications. Transfer learning (TL), under this context, is attracting increasing attention as it aims to harness information from dataset of other regions where labels are readily available. The central topic of concern is to homogenize the large disparities in terms of radiometry, geometry and scene contents through feature domain adaptation (DA) for cross-source and cross-region datasets. In this paper, we propose a novel DA method for unsupervised transferring learning, named Multi-Kernel Jointly Domain Matching (MKJDM). The proposed method by definition considers multiple kernels as opposed to the currently popular single-kernel methods when measuring the discrepancies between feature domain distributions. The single-kernel methods evaluate and minimize the distances of features between the source domain (dataset with readily and sufficiently available labels) and the target domain (dataset to be classified) through for example, Maximum Mean Discrepancy (MMD) metric, formed under a kernel function mapping. The Multi-kernel version (MK-MMD) maps the metric through different kernel functions and is able to encapsulate multiple aspects of distribution discrepancies and is therefore more capable when the feature distributions are being minimized. Additionally, MKJDM is able to align the marginal and class conditional distributions in the source and target domains and at the same time reweight instances through their representation. The proposed domain adaptation method is performed on both very high resolution (VHR) multispectral dataset and multi-modal remote sensing datasets (i.e. Orthophoto and Digital Surface Models). Our experiments are performed on cross-country datasets with respectively distinct land patterns, i.e. satellite dataset from typical cities in Argentina, Singapore and Haiti, in which we take Argentina dataset as the source domain data for training and the other two respectively serve as the target domain data. Using a simple statistical classifier trained on samples in the source domain, our experiments have shown that the overall classification accuracies are improved by 37.28% and 46.62% after processed by our MKJDM method. We have additionally compared our method with five state-of-the-art DA methods for transfer learning and the comparative experiments have shown that our method achieves the best performance among these.

(Corresponding author: Rongjun Qin)
Wei Liu is with the Department of Civil, Environment and Geodetic Engineering, The Ohio state University, Columbus, USA (email: liu.7052@osu.edu).
Rongjun Qin is with the Department of Civil, Environment and Geodetic Engineering and the Department of Electrical and Computer Engineering, The Ohio state University, Columbus, USA (email: qin.324@osu.edu)

***Index Terms*—Domain adaptation, transfer learning, remote sensing, image classification, distribution adaptation, unsupervised, stable index.**

I. Introduction

Classfication for characterizing land cover on remote sensing images embraces a wide range of applications in many fields, such as land cover / land use change, urban growth and disaster analysis [1, 2]. The often used supervised classification methods require a large amount of labeled data for training, while in practical applications such a labeled dataset is either not readily available or expensive to acquire. A possible way to address such an issue is to harness information from dataset of other regions where labels are readily available (source domain). Several factors such as viewing angle of sensors, sun elevation angle differences, different kinds of sensors, atmospheric changes, seasonal variations affecting the phenology of vegetation and land-cover patterns might cause a large difference between images acquired on different geographical areas or over the same area but at different time [3]. A non-transferred classifier, which is trained by the source dataset and directly applied on the testing dataset (target domain), may result in undesired classification results because of the large discrepancy between the source and target domains. Transfer learning (TL), which learns an adapted classifier for the target domain by utilizing the information from the existing available labeled data (source domain), has shown a promising prospect in computer vision field in recent years [3-5]. In our unsupervised transfer learning task, labeled data for training is only available in the source domain and the source and target domain are different in data distribution. In transfer learning field, domain adaptation (DA) is used for such a transfer learning task [6]. DA aims to transform the feature representations of the both domains, such that the joint probability distributions of the source domain and the target domain become more similar, thus leading to smaller discrepancies between the source and target domains. Though DA, transfer learning is able to greatly improve the performance and avoid expensive data labeling efforts at the same time.

The distance measurement method between the source and target domains is important for DA. Most existing DA methods use Maximum Mean Discrepancy (MMD) as the distance measurement for the statistical distance of distributions in a

projected space [6-9]. Reproducing Kernel Hilbert Space (RKHS) is a widely used projected space which corresponds to a kernel. After choosing the certain kernel of the RKHS, a linear mapping matrix, which is able to project the source and target datasets into a third space (RKHS), can be computed by minimizing MMD. The transferred source and target domains are expected to have similar feature distributions. The current DA methods based on single-kernel MMD find Gaussian kernel is the most effective [7, 9-12], while the bandwidth of kernel is often kept as a constant and this might not be optimal to account for a wide range of feature distributions presented in the source and target domains. To our best knowledge, a statistic based DA method that is able to perform optimal kernel selection does not exist.

Multi-Kernel MMD (MK-MMD) has been proposed as an optimal kernel selection method for MMD (see section 2.4 for details) [13]. MK-MMD is able to determine an optimal kernel formed by a weighted combination of multiple kernels based on the source and target datasets. Based on MK-MMD, we propose a novel domain adaptation method named Multi-Kernel Jointly Domain Matching (MKJDM) for transfer learning on remote sensing data. Our proposed MKJDM minimizes the discrepancy between the source and target domains through: 1) simultaneously aligning the marginal distributions (class-free feature distribution) and conditional distributions (per-class feature distribution) and 2) adaptively reweighting learning samples based on their contributions in domain adaptation, to further improve the classification accuracy.

Our experiments are performed both on very high resolution (VHR) multispectral datasets and multi-modal remote sensing datasets (including pixel-wise overlaid Orthophoto and Digital Surface Models (DSM)) with different scene contents. The rest of the sections are organized as follows: Section II introduces the related works in domain adaptation, and basics of the MK-MMD and MMD measures, as well as two methods mostly relevant to our work: 1) Transfer Component Analysis (TCA) and 2) Transfer Joint Matching (TJM). Section III describes the proposed MKJDM method in detail and Section IV presents the experimental dataset and feature extraction methods. The experimental results and the analysis are performed in Section V. Section VI concludes this paper by analyzing the pros and cons of the proposed method.

II. Related Work

In the field of computer vision, DA methods have been widely used in cases when training labels are not available, and have been applied to unsupervised learning and classification problems [12, 14-16]. The DA methods are often performed by minimizing the distance between the feature distributions of source and target domains. The distances can be defined using different approaches, for example, CORrelation ALignment (CORAL) [15] utilized the second-order statistics (covariance) of the source and target features as the distance between the two domains; Geodesic Flow Kernel (GFK) embedded source and target datasets in a Grassmann manifold and regarded the distance between two domains in such a manifold as the distance metrics [17]. Maximum Mean Discrepancy (MMD), among these metrics, is the most widely for DA [6, 9, 18]. Transfer component analysis (TCA), for example, is a classic DA method based on this metric that optimizes the summed MMD distances of all instances between the source and target domain. It has shown to be capable of generating new feature representations for both source and target domain with smaller distances. Applications of TCA using remote sensing datasets for transfer learning have demonstrated that this method is able to improve the accuracy of cross-scene classification to a notable level [19]. To further improve the performance of TCA, Long et al. [12] proposed a Transfer Joint Matching (TJM) method which incorporates an instance reweighting method into the optimization framework of TCA to reduce the effect of some training samples which are irrelevant with the same-class samples in the testing dataset. This type of methods takes the marginal distribution alignment as the goal for optimization, while this may perform well only when their feature domains have similar conditional distributions: For example, if the source dataset is from a residential area and the target dataset is from a downtown area, the proportion of buildings and vegetation is significantly different. Merely aligning marginal distribution might make the transferred classifier recognize some buildings as vegetation. In cases where the conditional distributions of the source and target domains are significantly different, minimizing their joint probability distributions distances can be particularly useful. By taking advantage of pseudo target label generated by non-transferred classifier, the conditional distributions of the feature domains can be built. With this idea, the Joint Distribution Adaptation (JDA) method proposed by Long et al [20] incorporated conditional distributions with the marginal distributions into the distance minimizing process and reported a better accuracy than aligning the marginal distributions alone.

MMD can be normally well estimated for two feature distributions mapped into a RKHS, while how to choose an appropriate RKHS is a key issue to success. The existing DA methods fix a RKHS for a particular task, but adaptively choosing optimal kernels can be crucial to further improve the DA accuracy. Gretton at al. [13] proposed MK-MMD for kernel choice which minimizes Type II error (the probability of two samples belonging to different distributions identified as the same), given an upper bound on Type I error (the probability of two samples belonging to the same distribution identified as). In [13], a family of kernels are linearly combined by different weights which can be computed adaptively by the samples in the source and target domains. This makes it possible to apply the optimal RKHS kernel in DA.

In this section, we first introduce the widely used and basic Maximum Mean Discrepancy (MMD) measurement (Section II-A), which is also the basic version of the more advanced Multi-Kernel Maximum Mean Discrepancy (MK-MMD, introduced in Section II-D) used in our proposed work. Based on the MK-MMD measurement, our method is built on two basic probability alignment methods 1): Transfer Component Analysis (TCA) and 2) Transfer Joint matching (TJM), and these two methods will be respectively introduced in Section

II-B and II-C.

A. Maximum Mean Discrepancy (MMD)

Maximum Mean Discrepancy (MMD) is accurate in finding samples that were generated from the same distribution [13]. Therefore, MMD can be seen as a measurement for the distance between probability distributions based on RKHS [21]. Let $\mathbf{X}_S \in \{\mathbf{x}_{S_i}\}_{i=1}^{n_s}$ and $\mathbf{X}_T \in \{\mathbf{x}_{T_j}\}_{j=1}^{n_t}$ be the feature sets of source domain $\mathcal{D}_s$ and target domain $\mathcal{D}_t$ over all the classes respectively, where n_s is the number of samples in the source domain and n_t is the number of samples in the target domain; the empirical estimate of the distance between $\mathbf{X}_S$ and $\mathbf{X}_T$ can be defined by MMD as,

$$\mathrm{Dist}(\mathbf{X}_S, \mathbf{X}_T) = \left\| \frac{1}{n_s}\sum_{i=1}^{n_s}\phi(\mathbf{x}_{S_i}) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi(\mathbf{x}_{T_j}) \right\|_{\mathcal{H}} \quad (1)$$

where $\mathcal{H}$ is a universal RKHS [32], and ϕ is a nonlinear transformation that maps the feature vector to $\mathcal{H}$. Instead of using an explicit function ϕ, Pan et al. [22] proposed to reformulate this as a kernel learning problem, in which the MMD distance between $\mathbf{X}_S$ and $\mathbf{X}_T$ can be written as [23]:

$$\mathrm{Dist}(\mathbf{X}_S, \mathbf{X}_T) = \mathrm{tr}(\mathbf{KM}) \quad (2)$$

where $\mathbf{K}=\phi([\mathbf{X}_S, \mathbf{X}_T])^{\mathrm{T}}\phi([\mathbf{X}_S, \mathbf{X}_T]) \in \mathbf{R}^{(n_s+n_t)\times(n_s+n_t)}$ is the kernel matrix in RKHS and $\mathbf{M} \in \mathbf{R}^{(n_s+n_t)\times(n_s+n_t)}$, with $M_{ij} = \frac{1}{n_s^2}$, if $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s$; $M_{ij} = \frac{1}{n_t^2}$ if $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t$; otherwise, $M_{ij} = \frac{-1}{n_s n_t}$. The readers may refer [23] for more details on this formulation.

B. Transfer Component Analysis (TCA)

Through aligning marginal distributions between feature datasets from source and target domain, TCA generates better feature representations across domains. The kernel matrix $\mathbf{K}$ in (2) can be decomposed as $\mathbf{K}=(\mathbf{K}\mathbf{K}^{-1/2})(\mathbf{K}^{-1/2}\mathbf{K})$. Consider a matrix $\widetilde{\mathbf{W}} \in \mathbf{R}^{(n_s+n_t)\times m}$ as a matrix which is able to transform the feature vectors to a m dimensional space. In general, $m \ll n_s+n_t$. The resulting kernel matrix is formulated as [7]

$$\widetilde{\mathbf{K}} = (\mathbf{K}\mathbf{K}^{-\frac{1}{2}}\widetilde{\mathbf{W}})(\widetilde{\mathbf{W}}\mathbf{K}^{-\frac{1}{2}}\mathbf{K}) \quad (3)$$

where the rank of $\widetilde{\mathbf{K}}$ is m, that means the dimensions of the space corresponding to $\widetilde{\mathbf{K}}$ is m. Then the distance between the mapped feature sets $\mathbf{X}_S^*$ and $\mathbf{X}_T^*$ can be written as

$$\mathrm{Dist}(\mathbf{X}_S^*, \mathbf{X}_T^*) = \mathrm{tr}(\widetilde{\mathbf{K}}\mathbf{M}) \quad (4)$$

Let $\mathbf{W}=\mathbf{K}^{-\frac{1}{2}}\widetilde{\mathbf{W}} \in \mathbf{R}^{(n_s+n_t)\times m}$. By the definition of $\widetilde{\mathbf{K}}$ in (3), the distance between $\mathbf{X}_S^*$ and $\mathbf{X}_T^*$ can be rewritten as:

$$\mathrm{Dist}(\mathbf{X}_S^*, \mathbf{X}_T^*) = \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}^{\mathrm{T}}\mathbf{W}) \quad (5)$$

where $\mathbf{W} \in \mathbf{R}^{(n_s+n_t)\times D}$ is the transformation matrix used to embed the features from their original space to RKHS. To minimize the criterion (5), a regularization term $\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{W})$ is used to control the complexity of $\mathbf{W}$. And then, the kernel learning problem is then written as:

$$\begin{aligned} \min \quad & \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}^{\mathrm{T}}\mathbf{W}) + \lambda \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}^{\mathrm{T}}\mathbf{W} = \mathbf{I} \end{aligned} \quad (6)$$

where λ is a regularization parameter, $\mathbf{I} \in \mathbf{R}^{m\times m}$ and $\mathbf{I}_{n_s+n_t} \in \mathbf{R}^{(n_s+n_t)\times(n_s+n_t)}$ are the identity matrices, $\mathbf{H} = \mathbf{I}_{n_s+n_t} - \frac{1}{n_s+n_t}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is the centering matrix, where $\mathbf{1} \in \mathbf{R}^{n_s+n_t}$ is the column vector with all ones. The constraint $\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$ is to avoid the trivial solution ($\mathbf{W}$= 0). The readers may refer [7] for more details about TCA.

C. Transfer Joint Matching (TJM)

Inspired by TCA, TJM also adopts MMD as the nonparametric distance to measure the difference of feature distributions in RKHS, but TJM reweights the instances belonging to the source domain at the same time. The schematic diagram of reweighting instance is shown in Fig. 1.

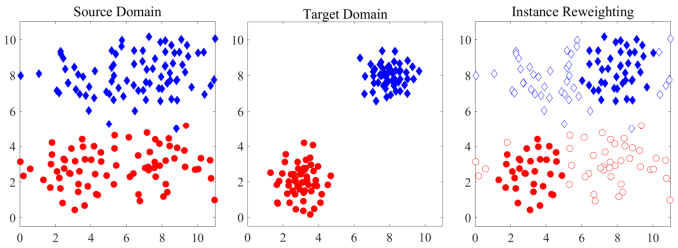


Figure 1. (a) instances in the source domain after aligning distribution (b) instances in the target domain after aligning distributions (c) instances in the source domain after aligning distribution and instance reweighting

In [12], the authors proposed to impose the $\ell 2,1$-norm structured sparsity regularization on the transformation matrix W, which introduces row-sparsity to the transformation matrix. Since each row of matrix W corresponds to an instance, the row-sparsity is able to essentially facilitate adaptive instance reweighting. The instance reweighting regularizer is defined as

$$\|\mathbf{W}_s\|_{2,1} + \|\mathbf{W}_t\|_F^2 \quad (7)$$

where $\mathbf{W}_s = (1{:}\,n_s, :)$ is the transformation matrix corresponding to the source instances, and $\mathbf{W}_t = \mathbf{W}(n_s + 1{:}\,n_s + n_t, :)$ is the transformation matrix corresponding to the instances in target domain. TJM only imposes $\ell_{2,1}$-norm regularization on source instances, since our aim is to reweight source instances by their relevance to the target instances. Then, the TJM optimization problem can be written as

$$\begin{aligned} \min \quad & \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}^{\mathrm{T}}\mathbf{W}) + \lambda(\|\mathbf{W}_s\|_{2,1} + \|\mathbf{W}_t\|_F^2) \\ \text{s.t.} \quad & \mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{H}\mathbf{K}^{\mathrm{T}}\mathbf{W} = \mathbf{I} \end{aligned} \quad (8)$$

The readers may refer [12] for more details about TJM. However, neither TCA nor TJM aligns the conditional distribution of the source and target domains. Although the distance between their marginal distributions are minimized, the distance between their joint distribution may not be minimized. In MKJDM, we will use pseudo labels to align the marginal and conditional distributions at the same time.

D. Multi-kernel Maximum Mean Discrepancy (MK-MMD)

Let $\mathcal{H}_u$ be the RKHS endowed with a Gaussian RBF (Radial Basis Function) kernel K_u whose bandwidth is σ_u. ϕ_u is the mapping function which can map the original data into $\mathcal{H}_u$. Combining a family of d kernels, the MK-MMD can be defined as:

$$H_{MK} = \mathrm{Dist}(\mathcal{D}_s, \mathcal{D}_t) = \sum_{u=1}^{d}\beta_u \left\| \frac{1}{n_s}\sum_{i=1}^{n_s}\phi_u(\mathbf{x}_{S_i}) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi_u(\mathbf{x}_{T_j}) \right\|_{\mathcal{H}_u} \quad (9)$$

where β_u is the weight of the u-th kernel k_u, $\sum_{u=1}^{d}\beta_u = 1$ ($\beta_u \geq 0$). We denote $\beta = (\beta_1, \beta_2, \cdots, \beta_d)^{\mathrm{T}} \in \mathbf{R}^{d\times 1}$ and

$\eta = (\eta_1, \eta_2, \cdots, \eta_d)^{\mathrm{T}} \in \mathbf{R}^{d\times 1}$ each η_u is the MMD via kernel k_u. With this notation,

$$\eta_{MK} = \beta^{\mathrm{T}}\eta \tag{10}$$

Following [13], we seek to learn the optimal kernel parameters β for MK-MMD by jointly maximizing the test power and minimizing the Type II error. Then the multi-kernel problem becomes equivalent to a convex quadratic program with a unique solution, given by

$$\begin{aligned} &\min \ \ \beta^{\mathrm{T}}(\mathbf{Q} + \lambda\mathbf{I})\beta \\ &\text{s.t. } \textstyle\sum_{u=1}^{d}\beta_u = 1 \ \ (\beta_u \geq 0) \end{aligned} \tag{11}$$

where $\lambda = 10^{-3}$ is a small regularization parameter to make the problem well-defined, $\mathbf{Q} \in \mathbf{R}^{d\times d}$ is the covariance matrix of $\mathbf{L} \in \mathbf{R}^{d\times \frac{n_s}{2}}$. $\mathbf{L}$ can be computed as follows,

$$\mathbf{L}(u, i) = k_u(\mathbf{x}_{S_{2i-1}}, \mathbf{x}_{S_{2i}}) + k_u(\mathbf{x}_{T_{2i-1}}, \mathbf{x}_{T_{2i}}) - k_u(\mathbf{x}_{S_{2i-1}}, \mathbf{x}_{T_{2i}}) - k_u(\mathbf{x}_{S_{2i}}, \mathbf{x}_{T_{2i-1}}) \tag{12}$$

$$k_u = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2{\sigma_u}^2}\right) \tag{13}$$

Our base kernel set $\{k_u\}_{u=1}^{d}$ contain d univariate kernels with fixed bandwidth. After solving the quadratic program in (11), the optimal linear combination for MK-MMD can be acquired. The readers may refer [13] for more details about MK-MMD.

III. The Proposed Multi-Kernel Jointly Domain Matching (MKJDM)

With the introduction of relevant works and basics, we in this section introduce our DA method that builds on TJM using MK-MMD, but with three main improvements to further improve the DA accuracy:

(1) TJM [12] merely aligns marginal distribution. Our MKJDM method aligns the marginal and conditional distributions simultaneous in a single optimization step.
(2) We use MK-MMD to measure the distance of the source and target domains, that can accommodate different aspects of distribution discrepancies.
(3) Because the reweighting instance method in MKJDM only reduces the weights of instances by row-sparsity, MKJDM only reweights the instances belonging to the classes appear in the classification result. Utilizing contextual information generated by pseudo labels is assumed to be more robust for conditional probability alignment.
(4) We apply the TCA result as the pseudo label set for estimate conditional distribution of the target domain.

To consider conditional distribution and marginal distribution at the same time, the MK-MMD of two domains in $\mathcal{H}_u$ can be expressed as:

$$\begin{aligned} \eta_u \approx &\left\| \frac{1}{n_s}\sum_{i=1}^{n_s}\phi_u(\mathbf{x}_{S_i}) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi_u(\mathbf{x}_{T_j}) \right\|_{\mathcal{H}_u}^2 \\ &+ \sum_{c=1}^{C}\left\| \frac{1}{n_c}\sum_{\mathbf{x}_{S_i}\in\mathcal{D}_s^{(c)}}\phi_u(\mathbf{x}_{S_i}) - \frac{1}{m_c}\sum_{\mathbf{x}_{T_j}\in\mathcal{D}_t^{(c)}}\phi(\mathbf{x}_{T_j}) \right\|_{\mathcal{H}_u}^2 \end{aligned} \tag{14}$$

where $c \in \{1, 2, \cdots, C\}$ is the distinct class label, n_c and m_c are the number of instances belonging to class c in source domain and target domain respectively. Similar to TJM and JDA, the distance between the empirical means of the two domains in (14) can be written as:

$$\mathrm{Dist}(\mathcal{D}_s, \mathcal{D}_t) = \mathrm{tr}(\mathbf{K}_u \textstyle\sum_{c=0}^{C}\mathbf{M}_c) \tag{15}$$

where $\mathbf{M}_0$ is corresponding to the marginal distribution distance measured by MK-MMD, with $\mathbf{M}_0(i,j) = \frac{1}{n_s^2}$, if $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s$; $\mathbf{M}_0(i,j) = \frac{1}{n_t^2}$ if $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t$; otherwise, $\mathbf{M}_0(i,j) = \frac{-1}{n_s n_t}$ and the MK-MMD conditional distance matrix $\mathbf{M}_c$ $(1 < c < C)$ can be computed as (16), $\mathbf{K}_u \in \mathbf{R}^{(n_s+n_t)\times(n_s+n_t)}$ is the kernel matrix of $\mathcal{H}_u$, $y(\mathbf{x}_i)$ is the label of $\mathbf{x}_i$, $\mathcal{D}_s^{(c)} = \{\mathbf{x}_i : \mathbf{x}_i \in \mathcal{D}_s \wedge y(\mathbf{x}_i) = c\}$, $\hat{y}(\mathbf{x}_j)$ is the pseudo label of $\mathbf{x}_j$, $\mathcal{D}_t^{(c)} = \{\mathbf{x}_j : \mathbf{x}_j \in \mathcal{D}_t \wedge \hat{y}(\mathbf{x}_j) = c\}$. In MKJDM, the pseudo label set is generated by a classifier transferred by TCA. Because TCA is independent from any label information, such a method is benefit for improving the accuracy of the pseudo label set.

$$(M_c)_{ij} = \begin{cases} \frac{1}{n_s^{(c)} n_s^{(c)}}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s^{(c)} \\ \frac{1}{n_t^{(c)} n_t^{(c)}}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t^{(c)} \\ \frac{-1}{n_s^{(c)} n_t^{(c)}}, & \mathbf{x}_i \in \mathcal{D}_t^{(c)}, \mathbf{x}_j \in \mathcal{D}_s^{(c)} \text{ or } \mathbf{x}_i \in \mathcal{D}_s^{(c)}, \mathbf{x}_j \in \mathcal{D}_t^{(c)} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

In this paper, we employ the Gaussian RBF kernel:

$$\mathbf{K}_u(i,j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2{\sigma_u}^2}\right) \tag{17}$$

where σ_u is the the kernel bandwidth. To acquire a family of kernels, we let σ_u vary from 0.025 to 2 with a step-size of 0.025. And then, the MK-MMD of the mapped two datasets can be written as

$$\eta_{MK} = \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{K}_M \textstyle\sum_{c=0}^{C}\mathbf{M}_c\mathbf{K}_M^{\mathrm{T}}\mathbf{W}) \tag{18}$$

where $\mathbf{K}_M = \sum_{u=1}^{d}\beta_u\mathbf{K}_u$ is the combined multi-kernel and β_u can be computed by solving (11). Similar to TJM, we impose the $\ell_{2,1}$-norm structured sparsity regularization on the transformation matrix for source domain $\mathbf{W}_s$ to control the complexity of the transformation matrix $\mathbf{W}$ and reweight the instances belonging to certain classes in source domain. The optimization can be formalized as:

$$\begin{aligned} &\min \ \ \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{K}_M \textstyle\sum_{c=0}^{C}\mathbf{M}_c\mathbf{K}_M^{\mathrm{T}}\mathbf{W}) + \lambda(\|\mathbf{W}_s\|_{2,1} + \|\mathbf{W}_t\|_F^2) \\ &\text{s.t. } \ \mathbf{W}^{\mathrm{T}}\mathbf{K}_M\mathbf{H}\mathbf{K}_M^{\mathrm{T}}\mathbf{W} = \mathbf{I} \end{aligned} \tag{19}$$

According to the constrained optimization theory, we denote $\mathbf{Z} = \mathrm{diag}(z_1, z_2, \cdots, z_k) \in \mathbf{R}^{k\times k}$ as the Lagrange multiplier and derive the Lagrange function for (19) as

$$\begin{aligned} &\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{K}_M \textstyle\sum_{c=0}^{C}\mathbf{M}_c\mathbf{K}_M^{\mathrm{T}}\mathbf{W}) + \mathrm{tr}((\mathbf{I} - \\ &\mathbf{W}^{\mathrm{T}}\mathbf{K}_M \textstyle\sum_{c=0}^{C}\mathbf{M}_c\mathbf{K}_M^{\mathrm{T}}\mathbf{W})\mathbf{Z}) + \lambda(\|\mathbf{W}_s\|_{2,1} + \|\mathbf{W}_t\|_F^2) \end{aligned} \tag{20}$$

Setting the derivatives of (20) w.r.t $\mathbf{W}$ to zero yields:

$$(\mathbf{K}_M \textstyle\sum_{c=0}^{C}\mathbf{M}_c\mathbf{K}_M^{\mathrm{T}} + \lambda\mathbf{G})\mathbf{W} = \mathbf{K}_M\mathbf{H}\mathbf{K}_M^{\mathrm{T}}\mathbf{W}\mathbf{Z} \tag{21}$$

$\|\mathbf{W}_s\|_{2,1}$ is a non-smooth function at zero, thus we compute its sub-gradient as $\frac{\partial(\|\mathbf{W}_s\|_{2,1} + \|\mathbf{W}_t\|_F^2)}{\partial\mathbf{W}} = 2\mathbf{G}\mathbf{W}$, where $\mathbf{G}$ is a diagonal sub-gradient matrix [12]:

$$\mathbf{G}(i,i) = \begin{cases} \frac{1}{2\|\mathbf{W}(i,:)\|_1}, & \mathbf{x}_i \in \mathcal{D}_s^{c}, \|\mathbf{W}(i,:)\|_1 \neq 0 \\ 0, & \mathbf{x}_i \in \mathcal{D}_s^{c}, \|\mathbf{W}(i,:)\|_1 = 0 \\ 1, & \text{otherwise} \end{cases} \tag{22}$$

where $\mathcal{D}_s^{c}$ is the source domain's instances belonging to the

classes appear in the pseudo labels acquired by a non-transferred classifier. Multiplying both sides in formula (21) on the left by $\mathbf{W}^{\mathrm{T}}$, we obtain

$$\mathbf{W}^{\mathrm{T}}(\mathbf{K}_M\sum_{c=0}^{C}\mathbf{M}_c\mathbf{K}_M^{\mathrm{T}}+\lambda\mathbf{G})\mathbf{W}=\mathbf{W}^{\mathrm{T}}\mathbf{K}_M\mathbf{H}\mathbf{K}_M^{\mathrm{T}}\mathbf{W}\mathbf{Z} \quad (23)$$

Substituting (23) into (20), the optimization becomes

$$\min \mathrm{tr}((\mathbf{W}^{\mathrm{T}}\mathbf{K}_M\mathbf{H}\mathbf{K}_M^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}(\lambda\mathbf{G}+\mathbf{K}_M\sum_{c=0}^{C}\mathbf{M}_c\mathbf{K}_M^{\mathrm{T}})\mathbf{W}) \quad (24)$$

or

$$\max \mathrm{tr}((\mathbf{W}^{\mathrm{T}}(\lambda\mathbf{G}+\mathbf{K}_M\sum_{c=0}^{C}\mathbf{M}_c\mathbf{K}_M^{\mathrm{T}})\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{K}_M\mathbf{H}\mathbf{K}_M^{\mathrm{T}}\mathbf{W}) \quad (25)$$

In kernel fisher discriminant (KFD) [24], the solution of $\boldsymbol{\alpha}$ for maximizing the $\frac{\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{M}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{N}\boldsymbol{\alpha}}$ is the leading eigenvectors of $\mathbf{N}^{-1}\mathbf{M}$. Therefore, the transformation matrix $\mathbf{W}\in\mathbf{R}^{(n_s+n_t)\times D}$ in (25) is the eigenvectors corresponding to the D leading eigenvalues of $(\lambda\mathbf{G}+\mathbf{K}_M\sum_{c=0}^{C}\mathbf{M}_c\mathbf{K}_M^{\mathrm{T}})^{-1}\mathbf{K}_M\mathbf{H}\mathbf{K}_M^{\mathrm{T}}$, where D is the dimensions of the data. By multiplying $\mathbf{W}^{\mathrm{T}}$ by the combination kernel matrix $\mathbf{K}_M$ on the right side, the original domains can be mapped into multi-kernel RKHS. Now, the new feature set $\mathbf{D}_{new}$ is acquired.

$$\mathbf{D}_{new}=\mathbf{W}^{\mathrm{T}}\mathbf{K}_M \quad (26)$$

The first n_s columns of $\mathbf{D}_{new}$ is the new feature representation in source domain and the last n_t columns of $\mathbf{D}_{new}$ is the new feature representation in target domain. And then we can classify the testing image by a simple statistical classifier using the new features.

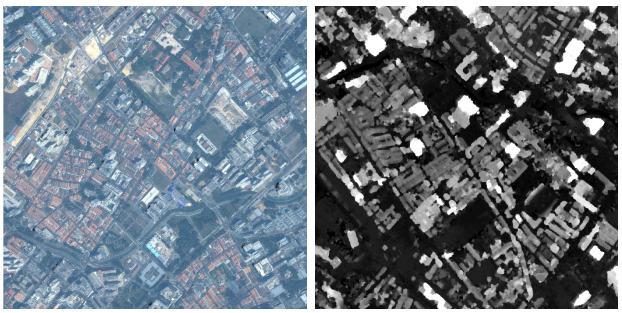
IV. Experiment Data and Processing

A. Experiment Dataset

In this paper, we used three urban areas from different continents with different land covers. Our first study area is a 3.5 km2 urban area in San Fernando (latitude 34° 27′ S and longitude of 58° 37′ W) which is a city in the Gran Buenos Aires, in Argentina, and capital of the San Fernando Partido (Fig. 2(a)). The second study area (Fig. 3) is Rochor which is a region from Singapore (latitude 1° 17′ N and longitude of 103° 50′ E) and the third study area (Fig. 4) is Port-au-Prince, Haiti (at latitude 18° 32′ N and longitude of 72° 20′ W), respectively. We use the Argentina dataset as the source domain and use the other two datasets as the target domains in this paper.

(a) (b)

Fig. 2. Dataset1: San Fernando, Argentina (the source domain)



(a) (b)

Fig. 3. Dateset2: Rochor, Singapore (the target domain in the first experiment)

(a) (b)

Fig. 4. Dataset3: Port-au-Prince, Haiti (the target domain in the second experiment)

B. Data description and processing

The experiments are conducted on three very high resolution (VHR) images (with or without DSM) obtained from Argentina, Singapore and Haiti respectively. In our experiments, we use the Argentina dataset as the source domain and use the other two datasets as the target domains. The Argentina dataset is a VHR orthophoto and DSM over San Fernando with a size of 6000 × 6000 pixels is computed by a hierarchical Semi-Global Matching algorithm (Hirschmuller et al. 2008) through RSP software [25-27]) using source satellite data from John's Hopkins University Applied Physics Lab's (JHUAPL) [28, 29] with the worldview-3 satellite which provides 8 spectral bands (red, red edge, coastal, blue, green, yellow, near infrared-1 and near infrared-2), with a 0.31m panchromatic resolution. The DSM of this area is shown in Fig. 2(b).

The Singapore dataset is an IKONOS Orthophoto (with DSM shown in Fig 3(b)) [30] which contains four 4 bands (blue, green, red, and near infrared). It was obtained over Singapore Rochor area with 1500 × 1500 pixels in 2002 with 1m resolution. And the Haiti dataset (Orthophoto + DSM) is obtained by the Geoeye-1 satellite over Port-au-Prince, the capital of Haiti. It was acquired in 2012 with 1.84m spatial resolution and contains four 4 bands (blue, green, red, and near infrared) with 2001 × 2001 pixels. We can see there are large discrepancies between the source domain (Argentina dataset) and the target domains (Singapore dataset and Haiti dataset), such as sensor, resolution, the number of bands, observation area and observation time. It is difficult for a classifier trained by the source domain to classify the images in target domains directly. However, our DA method is able to address this issue. Before implementing our idea, we need to pre-process the datasets to extract the original feature sets used to classify. In the rest part of this section, we introduce how to extract the

feature sets used in this paper.

By processing the DSM with morphology reconstruction algorithm, we can acquire the DSM features which make use of top-hat by reconstruction and erosion with different scales [31]. Let the normalized DSM image be J and J can be reconstructed as $B_{J,I}$ from a marker image I by finding the maximum of I, which is marked by J. I is derived from an erosion operation from J by a structuring element e. Considering two types of morphological top-hats, top-hat by reconstruction (THR) and top-hat by erosion (THE) can be simply defined as follows:

$$THR(J,e) = J - B_{J,I} \quad (27)$$

$$THE(J,e) = J - I \quad (28)$$

These two top-hats can compensate for each other. As the sizes of the urban objects vary a lot, we use a series of structuring elements $\{e_i\}_{i=1,2,\ldots,N}$ respectively to construct the multi-scale dual morphological top-hat profile (DMTHP) features:

$$DMTHP(J)_N = \{THR(J,e_1), THR(J,e_2), \cdots, THR(J,e_N), THE(J,e_1), THE(J,e_2), \cdots, THE(J,e_N)\} \quad (29)$$

In this paper, we use top-hat by reconstruction and erosion with 5 different scales: $\{e_i\}_{i=1,2,\ldots,N} = \{10, 30, 70, 120, 210\}$. The readers may refer [32] for more information regarding DMTHP. We use object-based classification method where mean shift method is used for segmentation. 1400 training samples are selected from the source domain. For accuracy verification purposes we have also labeled the target domain datasets. The feature vectors used in this paper contains 4 spectral features (red, green, blue and near infrared1) and ten DMTHP features. The other four bands in the Worldview-3 image are discarded in order to ensure these three datasets have the same features dimensions. All of the features are normalized to the range of [0, 1] and a random forest classifier is used to conduct all the experiments.

V. Results and Discussion

We compare MKJDM with some state-of-the-art domain adaptation methods: TCA [7], JDA [11], GFK [17], CORAL [15], TJM [12] in the experiments. We test MKJDM on both multi-model datasets (Orthophoto and DSM) and multi-spectral datasets (only have orthophoto). In addition, to understand the role of Multi-kernel MMD metric, we have implemented a single-kernel version of the MKJDM, denoted as JDM for comparison. The results show that our approach is able to improve classification accuracy in almost all cases (with/without DSM information). The statistical analysis of our algorithm in this paper uses multi-modal data as it achieves reasonably better improvements. Overall accuracy (OA) and Kappa index of agreement (KIA) [33] are used as metrics for accuracy evaluation.

A. Experiment-1: Singapore dataset as the target domain data

In this experiment, seven classes of interest have been interpreted, including buildings, roads, trees, bare land, grasses, ground and water. The DSM of Singapore dataset is shown in Fig 6(a) and the ideal classification result (from target to target: taking the labeled target samples to train the classifier, and note these were not used in the DA) is shown in Fig. 6(b).

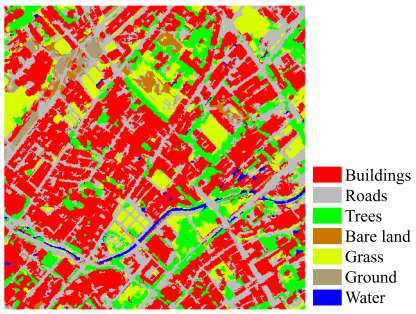
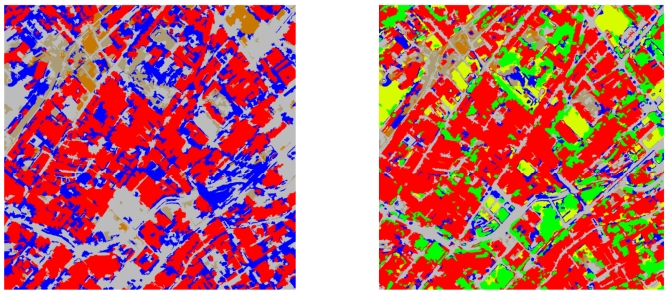


Fig. 5. The corresponding supervised classification result (from target to target).

From Fig. 2 and Fig. 3, we observe that there are many different accepts of the two scenarios such as the size of a typical building, the spectral reflection of the roofs, the width of the river, etc. Firstly, we adopt a family of RBF kernels whose bandwidths range between [0.025, 2] with a step-size of 0.025. Then, the final classification result acquired by MKJDM is shown in Fig 6(b). As a comparison, the classification result acquired by a non-transferred RF classifier is shown in Fig 6(a).



(a) Without Transfer learning
OA = 37.30% KIA = 24.94%

(b) After MKJDM
OA= 74.58% KIA=62.58%

Fig. 6. (a) the result acquired by a non-transferred RF classifier (b) the result acquired by a RF classifier transferred by MKJDM

In Fig 6(a), there are only 5 classes (buildings, waters, bare land, ground and road) in the classification result. Therefore, in MKJDM we only reweight the instances belonging to these 5 classes. From the comparison in Fig 6, we can see MKJDM is able to improve the classification performance greatly (The OA is improved from 37.30% to 74.58% and the KIA is improved from 24.94% to 62.58%).

In order to verify the effectiveness of MKJDM further, some state-of-the-art DA methods are compared with MKJDM in the Table I. Src-Tar represents the result acquired by a non-transferred classifier (classifier trained using the source labeled data and applied directly to the target dataset). Note that the JDM method is also proposed by us in this paper. Its result is able to be used to verify the validity of multi-kernel method.

TABLE I The results of different DA methods for Experiment 1

DA Method	Overall Accuracy (OA)	Kappa Index (KIA)
Src-Tar (with DSM)	0.3730	0.2494
TCA (with DSM)	0.5373	0.4328
JDA (with DSM)	0.6633	0.5536
GFK (with DSM)	0.3194	0.1655
CORAL (with DSM)	0.3703	0.2437
TJM (with DSM)	0.5388	0.4325
JDM (with DSM)	0.7031	0.6041
MKJDM (with DSM)	**0.7458**	**0.6258**
Src-Tar (image only)	0.3274	0.1977

MKJDM (image only)	0.5578	0.3936

We can see from Table I, almost all the DA methods is able to improve the classification effect. The OA of MKJDM is 37.28% higher than Src-Tar and 8% higher than the second best DA method. The comparison between MKJDM and JDM shows that the use of MKMMD is able to improve more than 4% of accuracy. The KIA of MKJDM is 7.22% higher than the second best DA method. In the case without using DSM features, the improvement from non-transfer learning (OA: 32.47%) to our proposed method (OA: 55.78%) is also significant. In a single kernel case, the bandwidth parameter associated with the RBF used in our method is 0.5.

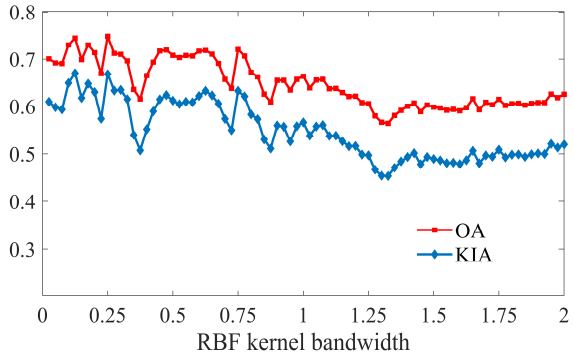


Fig. 7. The results acquired by JDM (single-kernel version of MKJDM) with different kernel bandwidths

Fig. 7 shows the performance varies with different kernel bandwidth. By using multi-kernel measurement (MK-MMD), the classification performance is much less sensitive as the reprojection to RKHS can be to appropriate kernels. In addition, MKJDM also reweights some instances in the source domain. We fix the range of kernel bandwidth as (0, 2] and change bandwidth step size in the kernel family. With the change of step-size, the number of the kernels in MKJDM also changes. Under different kernel family, we compare the results of MKJDM in three cases: 1) MKJDM with reweighting (our method); 2) MKJDM (without reweighting any instances) and 3) MKJDM (reweighting all the instances in source domain). Results are shown in Fig 9.

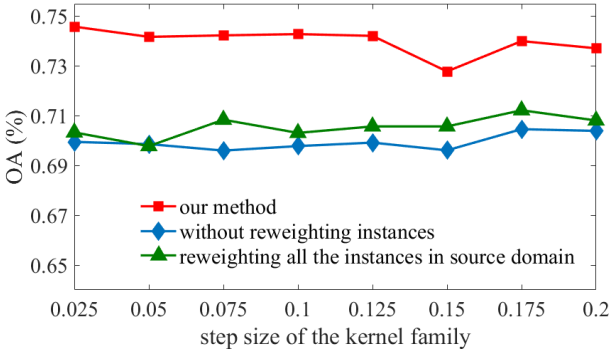


Fig. 8. The results acquired by reweighting all the instances, no instances and the instances belonging to certain classes in the source domain

In Fig. 8, we can see reweighting instances is able to improve the performance stably and reweighting part of samples are also better (about 3% higher in term of OA) than reweighting all the instances in source domain. From Fig. 8, we can see the step-size of bandwidths in the kernel family is not sensitive for MKJDM. Fig. 10 shows the comparison of class-wise original data distributions and class-wise transferred data (after MKJDM) distributions.

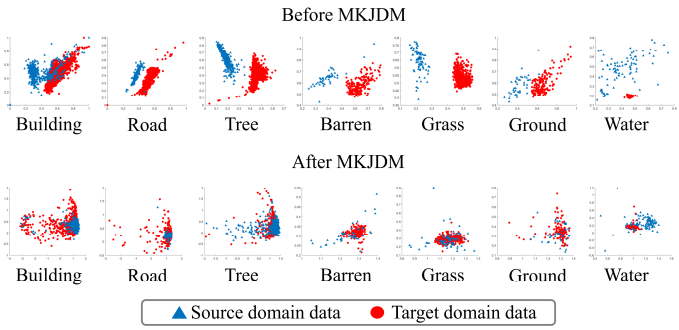


Fig. 9. The comparison of original data distributions and transferred data distributions

Considering joint distributions of the source and target domains, MKJDM maps the two datasets into a multi-kernel Hilbert space. After MKJDM, all the classes data distributions are dramatically closer. It is the fundamental reason why MKJDM can effectively improve the classification accuracy.

B. Experiment 2: Haiti dataset as the target domain data

In this dataset, six classes were interpreted as Experiment 1. The dataset is shown in Fig 4(a) and the ideal classification result (from target to target) shown in Fig. 10. In this experiment, six classes of interest have been interpreted, including buildings, roads, trees, bare land, grasses and ground. The DSM of the Haiti dataset is shown in Fig 4(b).

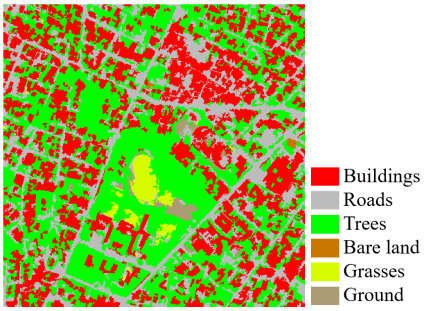
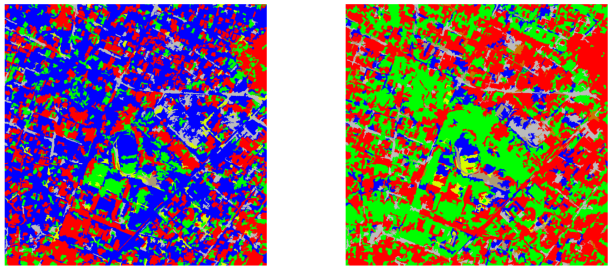


Fig. 10. The corresponding supervised classification result (from target to target).

By visual observation, there is a large discrepancy between the source (Argentina, southern hemisphere) and target (Haiti, middle-east) scene in terms of the object shapes, color spectrums and their distributions. As compared to the source domain, there is no water and the buildings are denser in Haiti dataset. In addition, the target image resolution is lower than source domain. We also adopt a family of RBF kernels whose bandwidths range between [0.025, 2] with a step-size of 0.025. The classification results acquired by MKJDM and a non-transferred classifier are shown in Fig 11.



(a) Without Transfer learning OA=29.90% KIA=14.79%

(b) After MKJDM OA= 76.52% KIA=62.37%

Fig. 11. (a) the result acquired by a non-transferred classifier (b) the result acquired by MKJDM

In this experiment, all the six classes are existed in Fig 11(a). Therefore, all the instances in source domain are reweighted in MKJDM. The comparative experiment results are shown in Table II.

TABLE II The results of different DA methods for Experiment 2

DA Method	Overall Accuracy (OA)	Final Kappa Index (KIA)
Src-Tar (with DSM)	0.2990	0.1479
TCA (with DSM)	0.7132	0.5410
JDA (with DSM)	0.7410	0.5811
GFK (with DSM)	0.3459	0.2167
CORAL (with DSM)	0.3022	0.1520
TJM (with DSM)	0.7184	0.5401
JDM (with DSM)	0.7482	0.5955
MKJDM (with DSM)	**0.7652**	**0.6237**
Src-Tar (image only)	0.2902	0.1435
MKJDM (image only)	0.5306	0.4983

All the DA methods are able to improve both OA and KIA of the classification result. Even if the OA of the Src-Tar is only 29.90%, MKJDM can also improve the classification result dramatically. The OA of MKJDM reaches 76.52% which is 46.62% higher than Src-Tar and 2.42% higher than the second best DA method. At the same time, the KIA of MKJDM reaches 62.37%, 47.58% higher than Src-Tar and 4.26% higher than the second best DA method. In addition, the OA is improved from 29.02% to 53.06% by our method in the case without DSM features. Using MKJDM, multi-model dataset has a better transferability across different remote sensing datasets. This is mainly because that height information is more stable across different dataset. Table II suggests that the MKJDM outperforms the single-kernel case JDM for about 2% in this dataset. We conducted similar analyses as per Experiment I on 1) MKMMD with and without reweighting (Fig. 13) and 2) feature distribution before and after alignment (Fig. 14), both suggests similar conclusions as Experiment I.

From experiments in this paper, with the bandwidth in such a range, JDM is able to achieve a relatively good performance. To see clearly, the results of JDM with different kernel bandwidths are shown in Fig 12. In Fig 12, the OA and KIA is the average result of JDM with different RBF kernels whose bandwidths range from 0.25 to 1.5 with a step-size of 0.025.

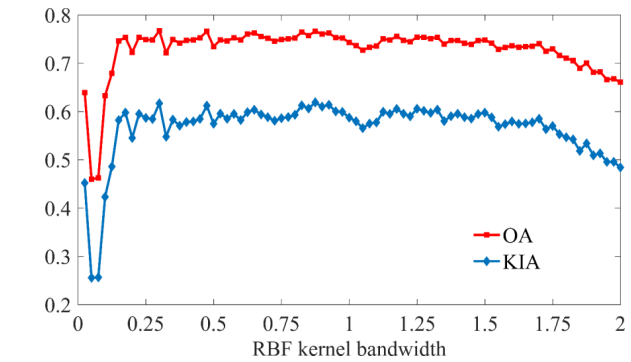


Fig. 12. The results acquired by JDM with different kernel bandwidths

From Fig 12, we can conclude the same conclusion as the experiment between Argentina and Singapore datasets. It demonstrates that multi-kernel MMD is better than single kernel MMD with any bandwidth.

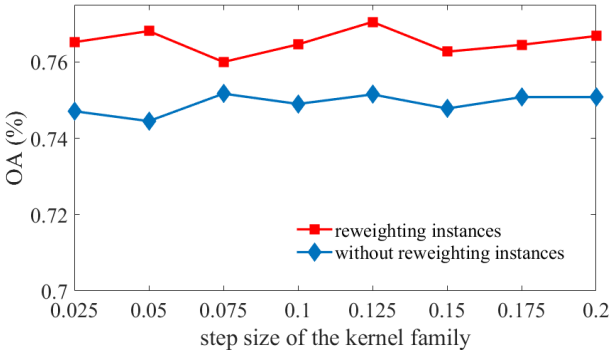


Fig. 13. The results acquired by MKJDM and MKJDM without reweighting instances

From Fig 13, we can see reweighting instances is able to improve the classification performance stably in the Haiti dataset. In this experiment, we also fix the range of kernel bandwidth as (0, 2]. The step of bandwidths in the kernel family is not sensitive for MKJDM. Fig. 14 shows the comparison of class-wise original data distributions and class-wise transferred data (after MKJDM) distributions. Although the Argentina dataset has seven classes and the Haiti dataset has only six classes, MKJDM is also able to make the transferred datasets be closer in each class.

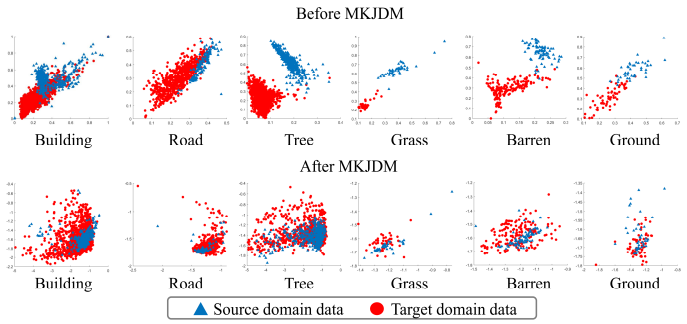


Fig. 14. The comparison of original data distributions and transferred data distributions

VI. Conclusion

In this paper, we propose a novel unsupervised domain

adaptation method named Multi-Kernel Jointly Domain Matching (MKJDM) method. The proposed method is applied to both multi-modal remote sensing dataset (Orthophoto + DSM) and multi-spectral dataset (with DSM). Based on our experiments, the proposed method has achieved satisfactory results and has a promising potential to significantly reduce the label cost in classifying remote sensing images. To align the distributions of the source and target feature domains, our method considers formulating both marginal distribution and conditional distribution under a multi-kernel MMD measure in a single optimization step. In addition, our proposed method reweights instances belonging to certain classes in the source domain to enhance the robustness of the feature distribution alignment. The solution of the formulated optimization delivers a transformation matrix which is used to map the feature sets of the source and target domains into Reproducing Kernel Hilbert Space (RKHS) can be computed, and the resulting feature distributions of the source and target domain are more statistically close. A simple statistical classifier using the transformed features for classification is able to acquire satisfactory results. As compare to other state-of-the-art DA methods, the proposed method obtained the best results in our experiments, and in the best case, has obtained 47% improvement as comparing to a non-transferred classifier. Our experiment datasets of cross-continental regions suggest the practical potentials in using unsupervised method for classification of remote sensing data through DA based transfer learning. In future works, we will consider including more stable features of the multi-modal to further improve the performance of the classification and reducing the computational complexity to make the algorithm can be used in large scale remote sensing images.

ACKNOWLEDGMENT

This work is partially supported by Sustainable Resilience Economy seed grant at the Ohio State University and ONR basic research grant (Award No. N000141712928). The author would like to acknowledge that the Singapore source dataset is from Singapore-ETH Center. The authors would like to thank the John's Hopkins University Applied Physics Lab and Digital Globe for making the Argentina Benchmark dataset available.

REFERENCES

[1] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 130, pp. 277-293, 2017.

[2] C. Gómez, J. C. White, and M. A. Wulder, "Optical remotely sensed time series data for land cover classification: A review," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 116, pp. 55-72, 2016.

[3] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 47, no. 7, p. 2218, 2009.

[4] A. Stumpf, N. Lachiche, J.-P. Malet, N. Kerle, and A. Puissant, "Active learning in the spatial domain for remote sensing image classification," *IEEE transactions on geoscience and remote sensing,* vol. 52, no. 5, pp. 2492-2507, 2014.

[5] C. Persello and L. Bruzzone, "Active learning for domain adaptation in the supervised classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 50, no. 11, pp. 4468-4483, 2012.

[6] 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.

[7] 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. 199-210, 2011.

[8] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," *arXiv preprint arXiv:1502.02791,* 2015.

[9] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen, "Balanced distribution adaptation for transfer learning," in *Data Mining (ICDM), 2017 IEEE International Conference on*, 2017, pp. 1129-1134: IEEE.

[10] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen, "Balanced distribution adaptation for transfer learning," in *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 1129-1134: IEEE.

[11] 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*, 2013, pp. 2200-2207: IEEE.

[12] 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.

[13] A. Gretton *et al.*, "Optimal kernel choice for large-scale two-sample tests," in *Advances in neural information processing systems*, 2012, pp. 1205-1213.

[14] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 1, no. 2, p. 4.

[15] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI*, 2016, vol. 6, no. 7, p. 8.

[16] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*, 2016, pp. 443-450: Springer.

[17] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2066-2073: IEEE.

[18] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," *ArXiv e-prints, abs/1704.01705,* 2017.

[19] 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.

[20] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2200-2207.

[21] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," *Journal of machine learning research,* vol. 2, no. Nov, pp. 67-93, 2001.

[22] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *AAAI*, 2008, vol. 8, pp. 677-682.

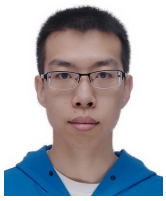
[23] L. Song, "Learning via Hilbert space embedding of distributions," 2008.

[24] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.*, 1999, pp. 41-48: Ieee.

[25] R. Qin, "Automated 3D recovery from very high resolution multi-view satellite images," presented at the ASPRS (IGTF) annual Conference, Baltimore, Maryland, USA, March 12-16, 2017.

[26] R. Qin, "Rpc stereo processor (rsp)–a software package for digital surface model and orthophoto generation from satellite stereo imagery," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* vol. 3, p. 77, 2016.

[27] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 30, no. 2, pp. 328-341, 2008.
[28] M. Bosch, Z. Kurtz, S. Hagstrom, and M. Brown, "A multiple view stereo benchmark for satellite imagery," in *Applied Imagery Pattern Recognition Workshop (AIPR), 2016 IEEE*, 2016, pp. 1-9: IEEE.
[29] M. Bosch, A. Leichtman, D. Chilcott, H. Goldberg, and M. Brown, "Metric Evaluation Pipeline for 3d Modeling of Urban Scenes," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* vol. 42, p. 239, 2017.
[30] R. Qin and W. Fang, "A hierarchical building detection method for very high resolution remotely sensed images combined with DSM using graph cut optimization," *Photogrammetric Engineering & Remote Sensing,* vol. 80, no. 9, pp. 873-883, 2014.
[31] R. Qin, J. Tian, and P. Reinartz, "3D change detection–approaches and applications," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 122, pp. 41-56, 2016.
[32] Q. Zhang, R. Qin, X. Huang, Y. Fang, and L. Liu, "Classification of ultra-high resolution orthophotos combined with DSM using a dual morphological top hat profile," *Remote Sensing,* vol. 7, no. 12, pp. 16422-16440, 2015.
[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement,* vol. 20, no. 1, pp. 37-46, 1960.



Wei Liu received the B.Eng. degree from Harbin Institute of Technology, Harbin, China, in 2016.
He is currently working toward the Ph.D. degree in the School of Electronics and Information Engineering, Harbin Institute of Technology. During the time of this work, he is a visiting Ph.D. student in The Ohio State University, Columbus, OH, USA. His research interests include transfer learning, computer vision and remote sensing image interpretation.

Mr. Liu serves as a Peer Reviewer for IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, IET Radar, Sonar & Navigation, IET image processing, etc.

Rongjun Qin received the B.S. degree in computational mathematics from Wuhan University, Wuhan, China, in 2009, and the M.S. and Ph.D. degree in photogrammetry and remote sensing from Wuhan University and ETH, Zurich in 2011 and 2015 respectively. He is currently a faculty member of the Department of Civil, Environmental and Geodetic Engineering, Department of Electrical and Computer Engineering at The Ohio State University, Columbus. His research interests include photogrammetric 3D reconstruction, remote sensing image classification, UAV images processing, image dense matching and change detection. His research seeks for computational solutions to various geometric and interpretation problems in an urban context using imaging sensors such as aerial/UAV imagery, LiDAR, and satellite multispectral/hyperspectral images. Prof. Qin is the author of RSP (RPC stereo processor) and MSP (multi-stereo processor) used for reconstructing 3D information from 2D images with high quality.

Prof. Qin is an associate editor for the Photogrammetric Engineering and Remote Sensing journal. He is also chairing the working group “Satellite Constellation for Remote Sensing” of International Society for Photogrammetry and Remote Sensing Commission. His awards include the first prize of Mathematical Modeling Contest and several other prominent scholarship awards. Prof. Qin serves as a reviewer for more than 15 international journals in the field photogrammetry and remote sensing.