Unsupervised Domain Adaptation via Discriminative Manifold Embedding and Alignment

You-Wei Luo,[1] Chuan-Xian Ren,[1]* Pengfei Ge,[1] Ke-Kun Huang,[2] Yu-Feng Yu[3]
[1]School of Mathematics, Sun Yat-Sen University, China
[2]School of Mathematics, JiaYing University, China
[3]Department of Statistics and Institute of Intelligent Finance, Guangzhou University, China
{luoyw28, gepengf}@mail2.sysu.edu.cn, rchuanx@mail.sysu.edu.cn
{kkcocoon, yuyufeng220}@163.com

Abstract

Unsupervised domain adaptation is effective in leveraging the rich information from the source domain to the unsupervised target domain. Though deep learning and adversarial strategy make an important breakthrough in the adaptability of features, there are two issues to be further explored. First, the hard-assigned pseudo labels on the target domain are risky to the intrinsic data structure. Second, the batch-wise training manner in deep learning limits the description of the global structure. In this paper, a Riemannian manifold learning framework is proposed to achieve transferability and discriminability consistently. As to the first problem, this method establishes a probabilistic discriminant criterion on the target domain via soft labels. Further, this criterion is extended to a global approximation scheme for the second issue; such approximation is also memory-saving. The manifold metric alignment is exploited to be compatible with the embedding space. A theoretical error bound is derived to facilitate the alignment. Extensive experiments have been conducted to investigate the proposal and results of the comparison study manifest the superiority of consistent manifold learning framework.

Introduction

In machine learning, large-scale datasets with annotations play a crucial role during the learning process. Convolutional Neural Networks (CNNs) achieves a significant advance in various tasks via a huge number of well-labeled samples (LeCun, Bengio, and Hinton 2015). Unfortunately, such data is actually prohibitive in many real-world scenarios. Applying the learned model in the new environment, i.e., the cross-domains scheme, will cause a significant degradation of recognition performance (Ren, Xu, and Yan 2018; Kim et al. 2019).

Unsupervised Domain Adaptation (UDA) is designed to deal with the shortage of labels by leveraging the rich labels and strong supervision from the source domain to the target domain, where the target domain has no access to the annotations. In fact, datasets composed of specifically exploratory factors and variants, such as background, style, illumination, camera views or resolution, often lead to the shifting distributions (i.e., the domain shift) (Shimodaira 2000; Moreno-Torres et al. 2012). According to the transfer theory established by Ben-David et al. (Ben-David et al. 2007; 2010), the primary task for cross-domain adaptation is to learn the discriminative feature representations while narrowing the discrepancy between domains.

Recent literature indicates that CNNs learn abstract representations with nonlinear transformations (Bengio, Courville, and Vincent 2013), which suppress the negative effects caused by variant explanatory factors in domain shift (Long et al. 2015). Pioneer works (Long et al. 2015; Ganin et al. 2016; Long et al. 2017; Sankaranarayanan et al. 2018) attempt to transfer the source classifier with sufficient supervision to the target domain by minimizing the discrepancy between the source and target domains. Though early adversarial confusion methods (Ganin et al. 2016; Sankaranarayanan et al. 2018; Pinheiro 2018), which is inspired by Generative Adversarial Nets (GANs) (Goodfellow et al. 2014), promise the generated features are domain-indistinguishable and form a well-aligned the marginal distributions, the conditional distributions are still not guaranteed (Long et al. 2018; Saito et al. 2018; Chen et al. 2019b).

Some latest methods achieve remarkable improvement in accuracy by employing the uncertainty information on the target domain, e.g., pseudo labels and soft labels (Long et al. 2018; Saito et al. 2018; Pinheiro 2018; Chen et al. 2019b). Though such information transduced from the source domain strengthens the discriminative ability of the target domain, there are still two points to be further explored. First, direct utilization of uncertainty information is risky and should be treated cautiously (Long et al. 2018), as the hard-assigned pseudo labels may change the intrinsic structure of data space (Ding and Fu 2019). Second, the batch-wise training in deep learning limits the capture of global information; thus models may be misled by some extreme local distributions.

In this paper, we develop a novel Riemannian manifold embedding and alignment framework. As the transferability and discriminability are both valuable (Chen et al. 2019b), the proposal reaches a consistent rule for these two properties. The main idea is to describe the domains by a sequence

*Corresponding Author.

of abstract manifolds. Enlightened by the successful application of soft labels for conditional coding and the multilayer embedding in (Long et al. 2017; 2018), a probabilistic discriminant criterion is proposed. Further, we extend this criterion to a global approximation scheme, which overcomes the dilemma of discriminant learning in batch-wise training. Inspired by previous attempts on manifold learning (Gong et al. 2012; Huang et al. 2017), we employ manifold metric to measure the domain discrepancy. The contributions are summarized as follows.

- To optimize the structure of the target domain and reduce the risk of uncertainty information simultaneously, a probabilistic discriminant criterion is developed. Specifically, an inter-class penalty supervised by ground-truth labels is built on the source domain; this penalty aims to construct a separable structure for classes. Then a probabilistic and truncated intra-class agreement is proposed on the target domain, which treats the classes of the source domain as anchors and acquires the inter-class separability transductively.
- Based on the above criterion, a global approximation scheme is extended. To capture the global structure, it combines the global information in the last epoch with data in the current batch. Since such approximation only requires access to the class-wise centers, it is actually memory-saving.
- The manifold alignment is developed to be compatible with the embedding discriminant space. It establishes a series of abstract descriptors (i.e. the basis) for original data, and aligns the domains by minimizing the discrepancy between the abstract descriptors, while most of noise are filtered. Further, a theoretical error bound is derived to facilitate the selection of components.

Related Work

Traditional UDA models usually focus on learning domain-invariant and discriminative features (Pan et al. 2010; Long et al. 2013). Based on the manifold assumption, plentiful metrics are developed to measure the distance between instances from source and target (Gong et al. 2012; Fernando et al. 2013). Deep learning methods enhance the transferability by exploring the representations that disentangle exploratory factors of variants hidden behind the data (Bengio, Courville, and Vincent 2013; Yosinski et al. 2014). The distribution alignment methods minimize the discrepancy of domains based on common statistics directly, e.g., the first-order statistic based on maximum mean discrepancies (MMD) (Sejdinovic et al. 2013; Long et al. 2015; Ren et al. 2019) and the second-order statistic based on covariance matrices (Sun, Feng, and Saenko 2016; Chen et al. 2019a). Inspired by the GANs (Goodfellow et al. 2014), lots of adversarial approaches with different purposes are developed. The most common usage of adversarial networks is to generate the representations that fool the domain discriminator, thus the distributions of domains are more similar (Ganin et al. 2016; Long et al. 2017; Pinheiro 2018). Domain-specific and Task-specific methods aim to tackle the issue of compact representations in high-level layers (Long et al. 2017; Saito et al. 2018; Kim et al. 2019; Lee et al. 2019; Ding and Fu 2019).

Though adversarial alignment generates well marginal distributions, the conditional distributions still need to be explored. Recent researches suggest that discriminability plays a crucial role in the formation of class distributions (i.e., the conditional distributions) (Long et al. 2018; Ding and Fu 2019; Chen et al. 2019b). Conditional Domain Adversarial Network (CDAN) (Long et al. 2018) encodes the target predictions into deep features and then models the joint distributions of features and labels. Batch Spectral Penalization (BSP) (Chen et al. 2019b) revisits the relation between transferability and discriminability via the largest singular value of batch features.

Multi-layer Remannian Manifold Embedding and Alignment

In this section, we propose the Discriminative Remannian Manifold Embedding and Alignment (DRMEA) framework.

Backgrounds and Motivations

In the classical manifold learning paradigm, to construct a compact and discriminative embedding space, a low-dimensional manifold is usually extracted from the originally high-dimensional data space. Specifically, the Riemannian manifold $\mathcal{M}$ usually consists of a certain object such as linear subspace, affine/convex hull, symmetric positive definite (SPD) matrix (Huang et al. 2017).

From the perspective of discriminative embedding, graph-based criterion (Yan et al. 2007) is widely adopted in the area of manifold learning and domain adaptation. Basically, those methods establish the instances-based connection graph or similarity graph to construct a separable space. Besides, as the primary assumption of domain adaptation is based on statistical distribution, the alignment based on covariance matrices, which lie on the Riemannian manifold, equips the domain with the manifold and statistical properties. Motivated by it, our work aims to embed the graph-based discriminant criterion to the target domain, which is represented as manifolds (i.e., the covariance matrices).

Given features $\mathbf{X} \in \mathbb{R}^{d\times n}$ and its mean vector $\bar{\mathbf{x}} \in \mathbb{R}^d$, where d denote the dimension of features and n represent the sample sizes. Denote by S the input space (e.g., Euclidean space, Hilbert space or Manifold space), the manifold learning aims to learning a specific nonlinear mapping

$$f: \ S \ \rightarrow \ \mathcal{M},$$

where $\mathcal{M}$ is the low-dimensional embedding manifold. Based on the SPD representation setting, the image of a given covariance matrix $\mathbf{C}(\mathbf{X}) = \frac{1}{n-1}(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T)(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T)^T \in \mathbb{R}^{d\times d}$ is a low-dimensional SPD matrix $\mathbf{C}' = f(\mathbf{C}) \in \mathbb{R}^{d'\times d'}$, where $\mathbf{1}_n$ is n-dimensional vector with all one elements and $(\cdot)^T$ is the transpose operation. Intuitively, learning of mapping function f can be deduced to find a nonlinear transformation

$$g: \ \mathbf{X} \ \mapsto g(\mathbf{X})$$

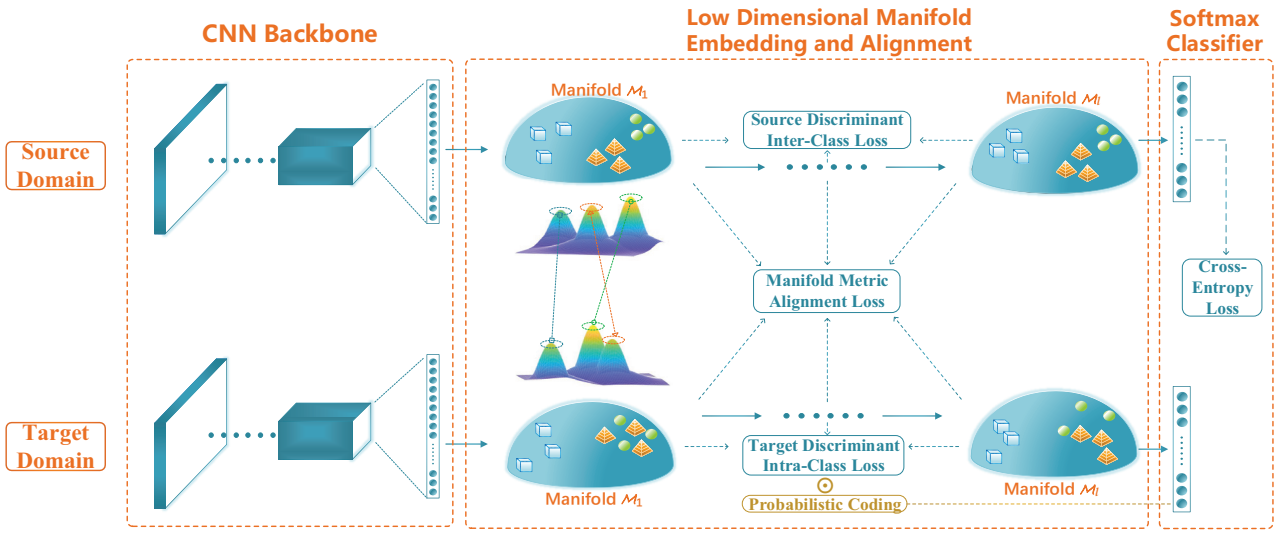

Figure 1: Overview of the proposed multilayer Riemannian manifolds embedding and alignment network. Stage 1: deep features based on CNNs. Stage 2: Riemannian manifold layers, where fully connected layers with proposed "weak" discriminant criterion and manifold metric domain alignment are employed to transfer the discriminative information.

and the image of mapping f can be approximated by the inner product of $g(\mathbf{X})$, i.e., $f(\mathbf{C}) \approx g(\mathbf{X})g(\mathbf{X})^T$.

For domain adaptation, the source and target domains can be taken as two Euclidean spaces, where the discriminative information is relatively inadequate. Thus the ideal manifolds are expected to be discriminative, representative and compact. Besides, the features distribution of domains, which is represented by manifolds, should be aligned with manifold metric for the better transfer of discriminative structure.

Low-Dimensional Manifold Layers

As previously stated, we aim to learn a nonlinear transformation g for the input features $\mathbf{X}$ directly. In this paper, CNNs are used to obtain such projection g. To explore the latent Riemannian representations of the Euclidean features (i.e., the deep features in stage 1), the output features of CNN backbone are sent into progressive low-dimensional manifold layers in the second stage. Since there is a naturally geometric difference between Euclidean Space and Riemannian Space, a multilayer scheme is adopted to reduce the dimension of features progressively.

Figure 1 shows the network architecture of the proposed method. Let Θ be the parameters of networks. The progressive Riemannian manifold layers $\{\mathcal{M}_i | i = 1, 2, \ldots, l\}$ are represented as a sequence of functions $\{g_i | i = 1, 2, \ldots, l\}$, and implemented on fully connection layers. In fact, the CNNs and Riemannian manifold layers are generalized and share by both two domains. It means that the common projections are explored to map two domains to a general low-dimensional space. Therefore, any manifold layers $\mathcal{M}_i$ should be equipped with the following properties:

- Discriminative Structure: To strengthen the discriminative power of manifold space, the intra-class samples are required to be compact, while the inter-class samples are separable, respectively.
- Consistent Structure: The source and target domains are aligned with manifold metric to match the manifold assumption. As a result, the domain discrepancy is represented as the distance between two submanifolds on $\mathcal{M}_i$, and then minimized based on the defined manifold metric (e.g., Grassmannian representations metric, Log-Euclidean metric and manifold principal angle similarity).

To reach the above goals, we propose to model the properties by losses $\mathcal{L}_{DS}$ and $\mathcal{L}_{AL}$, which will be detailed later. Then, the objective is formulated as following:

$$\min_{\Theta} \mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{DS} + \lambda_2 \mathcal{L}_{AL},$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss of classifier on source domain and $\{\lambda_1, \lambda_2\}$ are the penalty parameters.

Discriminative Structure Loss

In this section, we describe how to embed the discriminative structure into the manifold layers. The main idea is shown in Figure 2. Since there exists a distribution discrepancy between different domains (e.g., (a) in Figure 2), conventional discriminant criterion is too strong to satisfy in this case. To relax the constraint, our method only focuses on the inter-class separability of the source domain and the intra-class compactness of the target domain.

Without loss of generality, we only introduce the formulation of the loss terms in l-th Riemannian manifold layer $\mathcal{M}_l$. Let $\mathbf{H}_l^s \in \mathbb{R}^{d_l \times n_s}$ and $\mathbf{H}_l^t \in \mathbb{R}^{d_l \times n_t}$ be the feature matrices of $\mathcal{M}_l$. Since class centers of the source domain are used in both two loss terms, the source mean vector $\bar{\mathbf{h}}_l^s \in \mathbb{R}^{d_l}$ and source class-wise mean matrix $\bar{\mathbf{H}}_l^s \in \mathbb{R}^{d_l \times c}$ are computed, where c is the number of classes.

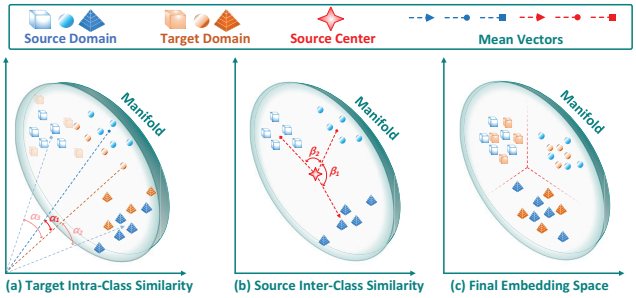

Figure 2: Illustration of the discriminative structure loss. (a) The target intra-class similarity constructs a compact space for the samples from same category. (b) The source inter-class similarity forms a separable space for source samples by finding a optimal rotations. (c) The final embedding space, where the target domain is discriminative.

Source Inter-Class Similarity Though the traditional inter-class discriminant criterion is applicable on the source domain, a nice geometric structure of the class distribution is actually not guaranteed under the distance metric. To this end, the similarity measurement is utilized here, which has also shown in Figure 2 (b). Rather than compute the similarities between class-wise centers and total center directly, we process the class-wise centers as following

$$\hat{\mathbf{H}}_l^s \triangleq \bar{\mathbf{H}}_l^s - \bar{\mathbf{h}}_l^s \mathbf{1}_c^T.$$

We call $\hat{\mathbf{H}}_l^s = [\hat{\mathbf{h}}_1^s, \hat{\mathbf{h}}_2^s, \ldots, \hat{\mathbf{h}}_c^s]$ the centralized class-wise means hereinafter. Further, if the columns of $\hat{\mathbf{H}}_l^s$ are normalized with ℓ_2 norm, the cosine similarity matrix is derived as $\mathbf{S}_{inter}^l = \hat{\mathbf{H}}_l^{s^T} \hat{\mathbf{H}}_l^s$. Because $\mathbf{S}_{inter}^l(i,j) = \hat{\mathbf{h}}_i^{s^T} \hat{\mathbf{h}}_j^s$ indicates the similarity between i-th class and j-th class, the diagonal elements are meaningless. Then the separable structure is reached by maximizing the dissimilarities between the centralized class-wise mean vectors. Equivalently, it can be achieved by minimizing the following inter-class loss:

$$\mathcal{L}_{inter}^l(\mathbf{H}_l^s) = \frac{2}{c(c-1)} \sum_{i<j} \mathbf{S}_{inter}^l(i,j). \tag{1}$$

Let us take Figure 2 (b) as an example. There is a 2-dimensional space with 3 classes. Let {1,2,3} denotes the labels of "Ball", "Pyramid" and "Cube", respectively. Under this situation, $\mathbf{S}_{inter}^l(1,2)$ and $\mathbf{S}_{inter}^l(1,3)$ are depicted as $\cos(\beta_1)$ and $\cos(\beta_2)$, respectively. According to the goal of Eq. (1) and ignoring the constraints, the optimal solution occurs at $\beta_1 = \beta_2 = \frac{2}{3}\pi$, and the minimal $\mathcal{L}_{inter}^l$ equals to $-\frac{1}{2}$ (which can also be seen as the lower bound of constrained scenarios).

Target Intra-Class Similarity On the other hand, since there are no labels on the target domain, the discriminant learning is facilitated by the soft labels (i.e., the output of softmax layer). Let $\mathbf{P}^t = [\mathbf{p}_1^t, \mathbf{p}_2^t, \ldots, \mathbf{p}_{n_t}^t] \in \mathbb{R}^{c \times n_t}$ be the softmax predictions of classifier layer. Since $\mathbf{P}^t$ can be regarded as the confidence or probability of classification, the predictions are used to weight the importance or confidence of the supervised information provided by soft labels. Similarly, assuming the columns of $\bar{\mathbf{H}}_l^s$ and $\mathbf{H}_l^t$ have unit length. The similarities under all classification cases can be written as $\mathbf{S}_{intra}^l = \bar{\mathbf{H}}_l^{s^T} \mathbf{H}_l^t$. It means that the source class-wise centers are utilized instead of the target. The main reasons can be summarized as follows: the inter-class structure learned from the source domain can be transduced to the target domain; the source class-wise centers computed from ground-truth labels are more reliable. Because there is so much uncertainty when pseudo labels are straightforwardly used on the target domain, we establish a probabilistic discriminative criterion to make the most of the information provided by soft labels. Intuitively, $\mathbf{P}^t$ is a natural choice for the probabilistically weighting model. Then the probabilistic intra-class loss is formalized as

$$\mathcal{L}_{intra}^l(\mathbf{H}_l^t, \mathbf{P}^t) = -\frac{1}{n_t c} \sum_{i=1}^{c} \sum_{j=1}^{n_t} \mathbf{P}^t(i,j) \mathbf{S}_{intra}^l(i,j). \tag{2}$$

However, there are much noise in $\mathbf{P}^t$, whose values are very small. Especially when the softmax classifier comes to converging, the columns of $\mathbf{P}^t$ tend to be the one-hot vectors. As truncation is a efficient way for denoising, we develop a Top-k preserving scheme for the truncated intra-class loss. Let $V_j = \{(i,j) | i = v_{1j}, v_{2j}, \ldots, v_{kj}\}$ be the index set of k-largest elements in $\mathbf{p}_j^t$, $j = 1, 2, \ldots, n_t$. Then a characteristic function like matrix is defined as

$$\chi(i,j) = \begin{cases} 1, & (i,j) \in V_j, \\ 0, & (i,j) \notin V_j. \end{cases}$$

Then, the intra-class loss is modified by the truncated matrix χ and written as

$$\mathcal{L}_{intra}^l(\mathbf{H}_l^t, \mathbf{P}^t) = -\frac{1}{n_t k} \sum_{i=1}^{c} \sum_{j=1}^{n_t} \chi(i,j) \mathbf{P}^t(i,j) \mathbf{S}_{intra}^l(i,j). \tag{3}$$

A simple illustration is also shown in Figure 2 (a). Based on the previous notations, $\mathbf{S}_{intra}^l(1,1)$, $\mathbf{S}_{intra}^l(1,2)$, $\mathbf{S}_{intra}^l(1,3)$ are computed as $\cos(\alpha_1)$, $\cos(\alpha_2)$ and $\cos(\alpha_3)$, respectively. Suppose the softmax output of the "Ball" sample in figure is $\mathbf{p}_1^t = [0.75, 0.15, 0.1]^T$. According to Eq. (2), all three similarities are taken into consideration, while $\cos(\alpha_2)$ and $\cos(\alpha_3)$ are noise. If we adopt the Top-2 strategy in Eq. (3), the perturbation from the "Cube" $\cos(\alpha_3)$ can be excluded.

In conclusion, the proposed two loss terms build a probabilistic discriminant criterion on the target domain. The ground-truth labels on the source domain provide a reliable separable structure directly, where the intra-class structure is unnecessary. Then the target samples are attached to the corresponding source class-wise center via soft labels. As shown in Figure 2 (c), the intra-class relationship on the source domain does not change much while the discriminative property of the target domain is satisfied. Finally, the discriminative structure loss is noted by

$$\mathcal{L}_{DS} = \sum_i (\mathcal{L}_{inter}^i + \mathcal{L}_{intra}^i),$$

Global Structure Learning For the batch scheme in deep models, it is hard to obtain the complete relation graph between the instances. The direct application of classical graph embedding may be misled by some extreme local distributions, which will result in a suboptimal solution.

Supposing that the geometry of manifolds does not change drastically after several updates, we can built some *anchors* in the whole data space to acquire the global information. In this work, we propose to fix the *anchors* in each batch iteration and update them after every epoch. Specifically, the *anchors*, i.e., $\bar{\mathbf{h}}_l^s$ in inter-class loss Eq. (1) and $\bar{\mathbf{H}}_l^s$ in intra-class loss Eq. (3), are computed from the last epoch. Note that the *anchors* are treated as constants in optimization. $\bar{\mathbf{H}}_l^s$ in Eq. (1) and $\mathbf{H}_l^t$ in Eq. (3) are obtained from batch data. The inter-class loss strongly supervised by source labels is imposed at the beginning. While the intra-class loss facilitated by soft label is equipped after a certain number of iterations/epoches.

Manifold Metric Alignment Loss

To satisfy the second property, i.e., Consistent Structure, a manifold metric alignment method is developed. As mentioned before, the covariance matrix is an important tool to represent a manifold $\mathcal{M}$. Therefore, the alignment based on covariance not only meets the requirements of manifold metric, but also reaches some nice statistical properties, such as distribution assumption.

Grassmannian Metric Let $\mathbf{C}_l^s$ and $\mathbf{C}_l^t$ be the covariance matrices of source and target domains computed from batch-wise features, respectively. Assume $\mathcal{M}_l^s$ and $\mathcal{M}_l^t$ are two submanifolds of $\mathcal{M}_l$, which are represented by their corresponding covariance matrices. Before the alignment process, these two submanifolds are partially overlapped, and our goal is to minimize the discrepancy under the metric of $\mathcal{M}_l$. In general, the manifold metric alignment loss of the l-th layer is expressed as

$$\mathcal{L}_{align}^l \triangleq dist(\mathcal{M}_l^s, \mathcal{M}_l^t) = d_{\mathcal{M}}(\mathbf{C}_l^s, \mathbf{C}_l^t), \tag{4}$$

where $d_{\mathcal{M}}(\cdot,\cdot)$ is the manifold metric to be determined.

Grassmannian manifold is a well-known type of Riemannian manifold. It is a projection subspace $\mathbb{R}^{d_l'}$ deduced from the originally high-dimensional space $\mathbb{R}^{d_l}$, $d_l' < d_l$. Thus the two submanifolds $\mathcal{M}_l^s$ and $\mathcal{M}_l^t$ lying on the Grassmannian manifold $\mathcal{M}_l$ are represented as two individual points. The distance between such two points is measured by the discrepancy between their projection orthogonal basis $\mathbf{U}_l^s$ and $\mathbf{U}_l^t$. Specifically, the orthogonal basis of such d_l'-dimensional Grassmannian manifold consists of d_l' dominant singular vectors with respect to its representation matrix. Thus $\mathbf{U}_l^s$ and $\mathbf{U}_l^t$ are two $d_l \times d_l'$ column-orthogonal matrices, which can be obtained from the Singular Value Decomposition (SVD) of covariance matrices $\mathbf{C}_l^s$ and $\mathbf{C}_l^t$, respectively. Finally, the Grassmannian distance is measured by

$$d_{\mathcal{M}}(\mathbf{C}_l^s, \mathbf{C}_l^t) = \frac{1}{d_l^2}\|\mathbf{U}_l^s\mathbf{U}_l^{s^T} - \mathbf{U}_l^t\mathbf{U}_l^{t^T}\|_F^2, \tag{5}$$

where $\|\cdot\|_F$ is the Frobenius norm. Thus the manifold metric alignment loss can be written as

$$\mathcal{L}_{AL} = \sum_i \mathcal{L}_{align}^i,$$

Error Bound of Grassmannian Metric As the dimension d_l' is needed in Grassmannian distance, we establish an theoretical error bound for it. Inspired by the previous works (Zwald and Blanchard 2006; Fernando et al. 2013), we shall denote the covariance of given distribution D by $\mathbf{C}$, and covariance drawn i.i.d. from D with sample size n by $\tilde{\mathbf{C}}$. Then Zwald et al. (Zwald and Blanchard 2006) give the following theorem.

Theorem 1. *(Zwald and Blanchard 2006) Let B be s.t. for any vector $\mathbf{x}$, $\|\mathbf{x}\| \leq B$, let $\mathbf{U}_{\mathbf{C}}^{d'}$ and $\mathbf{U}_{\tilde{\mathbf{C}}}^{d'}$ be the orthogonal projectors of the subspaces spanned by the first d' eigenvectors of $\mathbf{C}$ and $\tilde{\mathbf{C}}$, respectively. Let $\lambda_1 > \lambda_2 > \cdots > \lambda_{d'} > \lambda_{d'+1} \geq 0$ be the first $d'+1$ eigenvalues of $\mathbf{C}$, then for any $n \geq \left(\frac{4B}{\lambda_{d'}-\lambda_{d'+1}}\left(1+\sqrt{\frac{\ln(1/\delta)}{2}}\right)\right)^2$ with probability at least $1-\delta$ we have:*

$$\|\mathbf{U}_{\mathbf{C}}^{d'} - \mathbf{U}_{\tilde{\mathbf{C}}}^{d'}\| \leq \frac{4B}{\sqrt{n}\,(\lambda_{d'} - \lambda_{d'+1})}\left(1+\sqrt{\frac{\ln(1/\delta)}{2}}\right). \tag{6}$$

Above theorem shows the relation between the error and d'. Defining the right side of Eq. (6) as $\frac{E(\delta)}{\lambda_{d'}-\lambda_{d'+1}}$. To extend the inequality to the Grassmannian distance, we derive following lemma.

Lemma 2. *Based on the condition in Theorem 1, we have*

$$\|\mathbf{U}_{\mathbf{C}}^{d'}\mathbf{U}_{\mathbf{C}}^{d'^T} - \mathbf{U}_{\tilde{\mathbf{C}}}^{d'}\mathbf{U}_{\tilde{\mathbf{C}}}^{d'^T}\|_F \leq 2\sqrt{2}E(\delta)\frac{\sqrt{d'}}{\lambda_{d'} - \lambda_{d'+1}}$$

with probability at least $1-\delta$.

Based on Lemma 2, following theorem gives the error of $d_{\mathcal{M}}(\mathbf{C}^s, \mathbf{C}^t)$ with respect to its n samples approximation $d_{\tilde{\mathcal{M}}}(\tilde{\mathbf{C}}^s, \tilde{\mathbf{C}}^t)$.

Theorem 3. *Assuming the condition in Theorem 1 is specified by domains. Specifically, λ_i^s and λ_i^t denote the i-th largest eigenvalue of domain-specific covariance matrices $\mathbf{C}^s$ and $\mathbf{C}^t$, respectively. Denote by*

$$e(d') = \frac{\sqrt{d'}}{\lambda_{d'}^s - \lambda_{d'+1}^s} + \frac{\sqrt{d'}}{\lambda_{d'}^t - \lambda_{d'+1}^t}$$

the error index. Then the following error bound holds with probability at least $1-\delta$:

$$|d_{\mathcal{M}}(\mathbf{C}^s, \mathbf{C}^t) - d_{\tilde{\mathcal{M}}}(\tilde{\mathbf{C}}^s, \tilde{\mathbf{C}}^t)| \leq 2\sqrt{2}E(\delta)e(d').$$

Theorem 3 suggests that the upper bound of error is proportional to $e(d')$. It means that we should search the maximal gap between the continuous eigenvalues with the consideration of inflation factor $\sqrt{d'}$. Recall that in batch learning setting, the batch size b_s is usually smaller than d, thus d' only need to be searched in $\{1, 2, \ldots, b_s - 1\}$. The proofs are given in the Supplementary.

Table 1: Recognition Rates (%) on VisDA-2017 (ResNet-101) and Office-Home (ResNet-50).

VisDA-2017	Plane	bcycl	bus	car	horse	knife	mcyle	person	plant	sktbrd	train	truck	Mean
ResNet-101 (He et al. 2016)	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DAN (Long et al. 2015)	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
DANN (Ganin et al. 2016)	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MCD (Saito et al. 2018)	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
SimNet (Pinheiro 2018)	**94.3**	**82.3**	73.5	47.2	87.9	49.2	75.1	**79.7**	85.3	68.5	81.1	**50.3**	72.9
GTA (Sankaranarayanan et al. 2018)	-	-	-	-	-	-	-	-	-	-	-	-	77.1
CDAN (Long et al. 2018)	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.7
GPDA (Kim et al. 2019)	83.0	74.3	80.4	66.0	87.6	75.3	83.8	73.1	90.1	57.3	80.2	37.9	73.3
BSP+DANN (Chen et al. 2019b)	92.2	72.5	83.8	47.5	87.0	54.0	86.8	72.4	80.6	66.9	84.5	37.1	72.1
BSP+CDAN (Chen et al. 2019b)	92.4	61.0	81.0	57.5	89.0	80.6	90.1	77.0	84.2	**77.9**	82.1	38.4	75.9
DRMEA (No AL)	92.8	15.3	**86.7**	**86.3**	**93.8**	70.7	**95.2**	68.9	**95.8**	40.4	85.1	5.6	69.7
DRMEA (No DS)	90.2	66.5	70.2	65.8	79.8	81.8	84.7	70.1	82.0	46.5	**88.1**	27.7	71.1
DRMEA	92.1	75.0	78.9	75.5	91.2	**81.9**	89.0	77.2	93.3	77.4	84.8	35.1	**79.3**

Office-Home	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Mean
ResNet-50 (He et al. 2016)	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN (Long et al. 2015)	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN (Ganin et al. 2016)	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN (Long et al. 2017)	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN (Long et al. 2018)	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
CDAN+E (Long et al. 2018)	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
BSP+DANN (Chen et al. 2019b)	51.4	68.3	75.9	56.0	67.8	68.8	57.0	49.6	75.8	70.4	57.1	80.6	64.9
BSP+CDAN (Chen et al. 2019b)	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	**72.2**	**59.3**	**81.9**	66.3
DRMEA (No AL)	51.9	72.8	77.1	63.0	**72.0**	71.3	60.5	49.5	78.4	71.5	54.4	82.8	67.1
DRMEA (No DS)	51.2	72.4	**77.7**	63.0	71.4	71.4	58.6	44.6	**79.1**	71.1	53.4	81.5	66.3
DRMEA	**52.3**	**73.0**	77.3	**64.3**	**72.0**	**71.8**	**63.6**	**52.7**	78.5	72.0	57.7	81.6	**68.1**
	±0.4	±0.6	±0.3	±0.3	±0.7	±0.5	±0.6	±0.7	±0.2	±0.1	±0.6	±0.2	±0.2

Experiments and Comparative Analysis

In this section, three popular domain adaptation datasets are selected and the standard evaluation protocols are adopted.

Office-Home (Venkateswara et al. 2017) contains 4 domains, i.e., *Art* (**Ar**), *Clipart* (**Cl**), *Product* (**Pr**) and *Real-World* (**Rw**).

Image-CLEF-DA[1] consists of 4 domains. Following the previous protocol (Long et al. 2018), we conduct adaptation task between *Caltech* (**C**), *ImageNet* (**I**) and *Pascal* (**P**).

VisDA-2017 (Peng et al. 2017) is a large-scale visual domain adaptation challenge dataset. The **synthetic** data to **real-image** track is evaluated here.

Setup

Two layers Riemmanian manifold learning scheme is carried out in all experiments (i.e., $l = 2$), where the first layer (1024d) is activated by Leaky ReLU ($\alpha = 0.2$) and the second layer (512d) by Tanh. Adam Optimizer ($lr = 0.0002$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with batch size of 50 is utilized on Office-Home and Image-CLEF-DA datasets; the modified mini-batch SGD (Ganin et al. 2016) ($lr = 0.003$, momentum $= 0.9$, weight decay $= 5e-4$) with batch size of 32 is employed on VisDA-2017 challenge. The learning rate of CNN backbone layers is set as $0.1lr$. The hyperparameters are determined by try-and-error approach. Specifically, λ_1 and λ_2 are set as $1e1$ and $5e3$, respectively. The Top-1 scheme is adopted for the target intra-class loss in Eq. (3). For ablation study, the model without discriminative structure loss and manifold metric alignment loss are abbreviated as DRMEA (No DS) and DRMEA (No AL), respectively.

Results Analysis

[1]https://www.imageclef.org/2014/adaptation

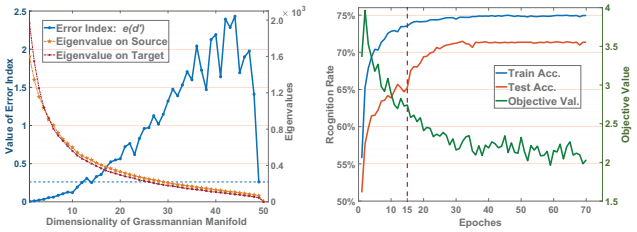

Figure 3: Left: Error and eigenvalue curves w.r.t. d'. Right: Recognition rate curves and loss curve.

Error Bound of Grassmannian Distance The numerical simulation is conducted on Image-CLEF-DA dataset to explore the minimal error index $e(d')$. As a fact that the eigenvalues always decrease rapidly at the beginning and enter into a flat state, the error bounds of dimensionality d' located in the flatten area are too high to assess. As shown in Figure 3, the trend of eigenvalues is consistent with the description. Though the dramatic decrease in the beginning stage results in a lower error, the information in that area is unconvincing and insufficient to support the measurement of manifolds. Since there is a natural gap between the (b_s-1)-th and b_s-th dominant eigenvalues, $e(b_s-1)$ is smaller than most of other errors. We highlight the error index of $e(b_s-1)$ by blue dash line, and observe only errors of $d' = \{1, 2, \ldots, 12, 14\}$ are lower than $e(b_s-1)$. Empirically, the dimensionality of Grassmannian manifold d' is set as b_s-1 hereinafter.

Convergence The convergence cures on Office-31 **A**→**W** adaptation task are displayed in Figure 3. It the beginning, the objective loss value decreases quickly and the recognition rate tends to enter a stable region in the epoch 10-15. However, the intra-class constraint is imposed after 15 epoches, which further activates the learning of discrimina-

Table 2: Recognition Rates (%) on Image-CLEF-DA (ResNet-50).

Image-CLEF-DA	I→P	P→I	I→C	C→I	C→P	P→C	Mean
ResNet-50 (He et al. 2016)	74.8 ± 0.3	83.9 ± 0.1	91.5 ± 0.3	78.0 ± 0.2	65.5 ± 0.3	91.2 ± 0.3	80.7
DAN (Long et al. 2015)	74.5 ± 0.4	82.2 ± 0.2	92.8 ± 0.2	86.3 ± 0.4	69.2 ± 0.4	89.8 ± 0.4	82.5
DANN (Ganin et al. 2016)	75.0 ± 0.3	86.0 ± 0.3	96.2 ± 0.4	87.0 ± 0.5	74.3 ± 0.5	91.5 ± 0.6	85.0
JAN (Long et al. 2017)	76.8 ± 0.4	88.0 ± 0.2	94.7 ± 0.2	89.5 ± 0.3	74.2 ± 0.3	91.7 ± 0.3	85.8
CDAN (Long et al. 2018)	76.7 ± 0.3	90.6 ± 0.3	97.0 ± 0.4	90.5 ± 0.4	74.5 ± 0.3	93.5 ± 0.4	87.1
CDAN+E (Long et al. 2018)	77.7 ± 0.3	90.7 ± 0.2	**97.7** ± 0.3	**91.3** ± 0.3	74.2 ± 0.2	94.3 ± 0.3	87.7
DRMEA (No AL)	78.0 ± 0.1	91.1 ± 0.1	95.6 ± 0.2	88.7 ± 0.3	74.8 ± 0.1	94.8 ± 0.2	87.3
DRMEA (No DS)	78.9 ± 0.1	90.5 ± 0.2	94.0 ± 0.1	87.8 ± 0.1	76.7 ± 0.2	93.0 ± 0.1	86.8
DRMEA	**80.7** ± 0.1	**92.5** ± 0.1	97.2 ± 0.1	90.5 ± 0.1	**77.7** ± 0.2	**96.2** ± 0.2	**89.1**

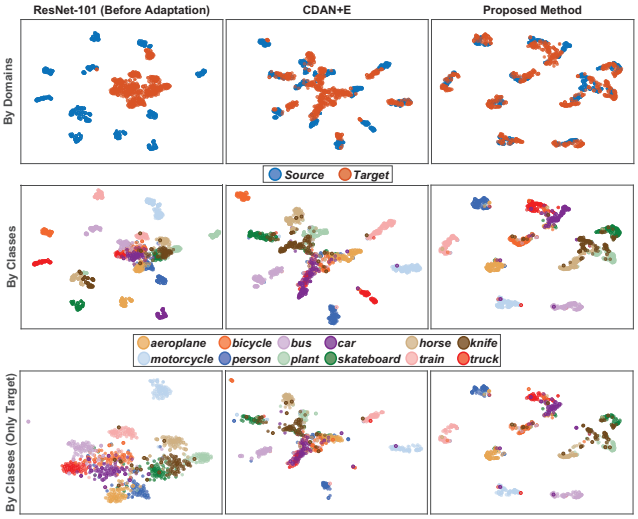

Figure 4: Visualization of the feature space on VisDA-2017 dataset. Rows are colored by domains or classes.

tive structure. Thus the second ascent of accuracy on target domain occurred after 15 epoches, which leads to the continuous improvement of recognition rate and alleviates the over-fitting the on the source domain.

Comparison Several state-of-the-art UDA approaches are selected and shown in Table 1-2. The experimental result on Visda-2017 dataset is shown in the top of Table 1. We observe that DRMEA outperforms others by a large margin in average accuracy from the result. Performance on Office-Home dataset is provided at the bottom of Table 1, the proposed method improves the accuracy to 68.1% and obtains the highest accuracy in most of the adaptation tasks. Results on Image-CLEF-DA dataset are provided in Table 2. As the discrepancy between the source and target domains on Image-CLEF-DA dataset is relatively smaller than others, a more discriminative model is essential to the improvement of recognition accuracy. DRMEA encodes the discriminant criterion and alignment constraint simultaneously, thus it outperforms other methods by 1.4% at least.

The ablation results also prove that the whole Riemannian manifold learning framework effect when both loss terms are equipped. As the discriminative structure loss provides a separable structure and manifold metric alignment loss bridges the distribution discrepancy between the source and target domains based on Grassmannian distance, both two losses are important.

Visualization Figure 4 shows the 2-D representation spaces obtained from t-SNE (Maaten and Hinton 2008) algorithm on VisDA-2017 dataset. CDAN+E shortens the distance between source and target by using adversarial alignment. A part of classes has been dragged away from the center, e.g., *plant, car, horse, aeroplane* and *bicycle*. In the third column, our method further optimizes the structure of the representation space. The categories are aligned better than ResNet-101 and CDAN+E, which leads to more compact target space.

Conclusion

In this paper, we develop a Riemannian manifold embedding and alignment framework for UDA, where the transferability and discriminability are reached consistently. To optimize the structure of the target domain, the soft labels are encoded into the discriminant criterion probabilistically and transductively. Then a globally discriminative structure is approximated via a memory-saving manner. A theoretical error bound is derived, which is guaranteed to find an appropriate dimension for manifolds during the alignment. Numerical simulation and extensive comparisons demonstrate the effectiveness of the derived theorem and proposed method. How to further reduce dependence of our proposal on temporal target predictions is our future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants 61976229, 61976104, 61906046, 61572536, 11631015 and U1611265.

References

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, 137–144.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine Learning* 79(1-2):151–175.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8):1798–1828.

Chen, C.; Chen, Z.; Jiang, B.; and Jin, X. 2019a. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *AAAI Conference on Artificial Intelligence*.

Chen, X.; Wang, S.; Long, M.; and Wang, J. 2019b. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, 1081–1090.

Ding, Z., and Fu, Y. 2019. Deep transfer low-rank coding for cross-domain learning. *IEEE Transactions on Neural Networks and Learning Systems* 30(6):1768–1779.

Fernando, B.; Habrard, A.; Sebban, M.; and Tuytelaars, T. 2013. Unsupervised visual domain adaptation using subspace alignment. In *IEEE International Conference on Computer Vision*, 2960–2967.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17(1):2096–2030.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2066–2073.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, Z.; Wang, R.; Shan, S.; Van Gool, L.; and Chen, X. 2017. Cross euclidean-to-riemannian metric learning with application to face recognition from video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(12):2827–2840.

Kim, M.; Sahu, P.; Gholami, B.; and Pavlovic, V. 2019. Unsupervised visual domain adaptation: A deep max-margin gaussian process approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4380–4390.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553):436.

Lee, C.-Y.; Batra, T.; Baig, M. H.; and Ulbricht, D. 2019. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10285–10295.

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2013. Transfer feature learning with joint distribution adaptation. In *IEEE International Conference on Computer Vision*, 2200–2207.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 97–105.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, volume 70, 2208–2217.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, 1640–1650.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Moreno-Torres, J. G.; Raeder, T.; Alaiz-RodríGuez, R.; Chawla, N. V.; and Herrera, F. 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45(1):521–530.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2010. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2):199–210.

Peng, X.; Usman, B.; Kaushik, N.; Hoffman, J.; Wang, D.; and Saenko, K. 2017. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*.

Pinheiro, P. O. 2018. Unsupervised domain adaptation with similarity learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8004–8013.

Ren, C.; Ge, P.; Dai, D.; and Yan, H. 2019. Learning kernel for conditional moment-matching discrepancy-based image classification. *IEEE Transactions on Cybernetics*.

Ren, C.; Xu, X.; and Yan, H. 2018. Generalized conditional domain adaptation: A causal perspective with low-rank translators. *IEEE Transactions on Cybernetics*.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3723–3732.

Sankaranarayanan, S.; Balaji, Y.; Castillo, C. D.; and Chellappa, R. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8503–8512.

Sejdinovic, D.; Sriperumbudur, B.; Gretton, A.; Fukumizu, K.; et al. 2013. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics* 41(5):2263–2291.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90(2):227–244.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of frustratingly easy domain adaptation. In *AAAI Conference on Artificial Intelligence*.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5018–5027.

Yan, S.; Xu, D.; Zhang, B.; Zhang, H.-J.; Yang, Q.; and Lin, S. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1):40–51.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, 3320–3328.

Zwald, L., and Blanchard, G. 2006. On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems*, 1649–1656.