24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Adaptive Local Neighbors for Transfer Discriminative Feature Learning

Wei Wang[1] and **Zhihui Wang**[1*] and **Haojie Li**[1] and **Juan Zhou**[2] and **Zhengming Ding**[3]

Abstract. In Domain Adaptation (DA), how to reduce the distributional differences across domains and preserve the data structures are two critical issues to obtain domain-invariant features. Existing DA methods either preserve the Local Manifold Structure (LMS) or the Global Discriminative Consistency (GDC), while fail to take those two metrics into account simultaneously. Therefore, the extracted features are either short of discriminative ability or sensitive to the multimodally distributed data. Moreover, the local neighbored relationships among data points are mostly established in original data space, which is unreliable, especially for data with large noises. Therefore, this paper proposes a novel DA approach, i.e., Adaptive Local Neighbors for Transfer Discriminative Feature Learning, to leverage LMS and GDC into a unified transfer feature learning model, where we only focus on the GDC between the local neighbors, so that the extracted features are more discriminative and robust to the multimodally distributed data. Moreover, the data points' local neighbors are revealed adaptively in the learned subspace so that it is insensitive to the data noises. Compared with the state-of-the-art methods, the proposed approach achieves higher performance for different cross-domain image classification tasks, especially 3.0% improved for Office10+Caltech10 dataset.

1 Introduction

Transfer Learning (TL) targets at transferring knowledge from a related source domain to the target domain, where rich source labels are available while there are few or no labels in the target domain. Recently, TL has made remarkable applications in cross-domain image classification [30, 11, 19], person re-identification [38, 4, 1], semantic segmentation [17, 34, 36], etc. Different from traditional machine learning, TL usually assumes that the source and target data are sampled from different distributions [5, 14]. Therefore, a major challenge of TL is to reduce the distributional differences across domains [28, 7]. One effective technique in TL is to learn the domain-invariant features, which specifically integrates with certain dimensionality reduction methods for discovering a projection to map different domains data into a common feature subspace, where the distributional differences across domains could be minimized [25].

Apart from the distributional alignment, it is also essential to preserve data structures hidden in original space. Recently, Transfer Discriminative Feature Learning (TDFL) has aroused great interest in either shallow or deep works, since it could enable the domain-invariant features more discriminative. For example, VDL [12], VDA [32], JGSA [39], SCA [6] and DICD [15] were proposed to maximize the inter-class dispersion and minimize the intra-class scatter, so that the class discriminative consistency could be respected. Long et al. revealed an unexpected deterioration of the discriminability while learning transferable features adversarially, and proposed a general approach to boost the feature discriminability [3]. However, they usually assume that the input data from the same class obey the unimodal distribution globally, and fail to deal with the multimodally distributed data since this kind of constraint (Global Discriminative Consistency, GDC) is too strong. As shown in Fig. 1, most data points from one class are located in more than one group, thus it will damage the local data structure if we enable those data points closer by force, then the feature transferability is degraded since it might also interfere the shared projection learning.

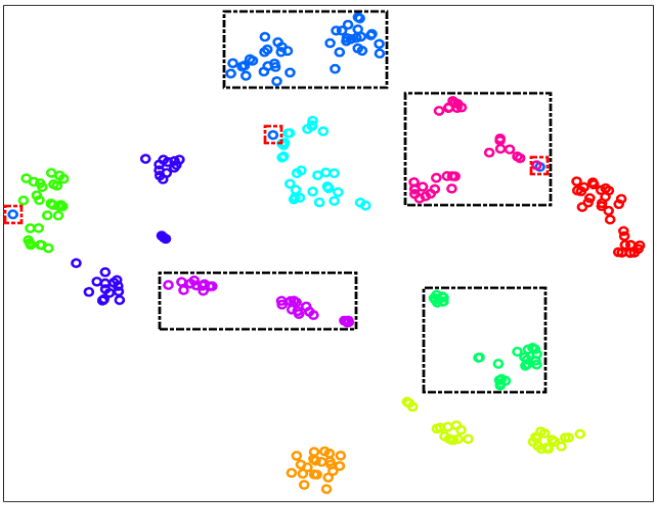

Figure 1. The images of 10 categories in Webcam dataset (denoted by different colors). (1) In real world applications, the data may be multimodally distributed, i.e., the data points of some classes are located in more than one group (inside the black dashed boxes); (2) Some noises inside the red dashed boxes, are fairly closer to intrinsically dissimilar points.

Therefore, it is imperative to capture the Local Manifold Structure (LMS) hidden in original space, since it does not require the data points from the same class to draw closer globally, and it is insensitive to the multimodally distributed data. In this regard, TSC [22], GTL [24], TRSC-GJDA [42], ARTL [23] and MEDA [35] were proposed to respect the LMS so that the embedded representations of two data points are closer if they are k-nearest neighbors to each other. Although LMS could compensate for the multimodally dis-

[1] DUT-RU International School of Information Science & Engineering, Dalian University of Technology, Dalian 116000, China. email: WWLoveTransfer@mail.dlut.edu.cn, {zhwang, hjli}@dlut.edu.cn
[2] College of Criminal Investigation, Southwest University of Political Science and Law, Chongqing 401120, China. email: zhoujuan@swupl.edu.cn
[3] Department of Computer, Information and Technology, Purdue School of Engineering and Technology, Indiana University-Purdue University Indianapolis, Indianapolis IN 46202, USA. email: zd2@iu.edu.cn

tributed data, the extracted features are short of discriminative ability. Intuitively, we should take both advantages of the LMS and GDC to refine more effective domain-invariant features, while the LMS-wise DA methods usually find the points' neighbors based on their distances in original data space, which is unreliable [27]. As shown in Fig. 1, closer points in the original space may be intrinsically dissimilar, especially for data with large noises (the data points inside the red dashed boxes). Moreover, it is nontrivial to obtain a unified framework, and optimize each variable quantity effectively, so that both LMS and GDC are respected elegantly.

Different from previous work, this paper proposes to conduct the transfer discriminative feature learning based on the adaptive local neighbors. Specifically, we construct a similarity weight matrix for each class, and the weight between two data points is larger if they are closer than anyone else, where all data points are from the same class. With this local information, the strong constraint of global discriminative consistency could be relaxed, since only the similar points from the same class are required to drawn closer. Therefore, the model could enable the extracted features more discriminative, and respect the LMS to deal with the multimodally distributed data simultaneously. On the other hand, the similarity weight matrix is exploited adaptively in the learned subspace instead of the original data space. Therefore, the effect of data noises in original space could be further mitigated. In order to verify the proposed approach is insensitive to data noises, we randomly corrupt the input features on the Office10+Caltach10 dataset, then compare the performance of ALN-TDFL with TDFL. Finally, the shared projection, and the points' local neighbors are optimized effectively by an efficient optimization strategy. The main contributions of our work are three-folds:

- We take both advantages of LMS and GDC metrics, so that the model could deal with the challenge of multimodally distributed data, and enable the extracted features more discriminative.
- We further mitigate the effect of data noises, and adaptively exploit the points' neighbors in the desired subspace so that the local manifold structure could be revealed correctly.
- Finally, we develop an efficient optimization strategy to learn the shared projection, and the points' local neighbors automatically.

2 Related Work

Existing feature-wise DA methods are devoted to exploit a shared feature subspace, where the distributional differences across domains are reduced and the data properties hidden in original space preserved [6, 12, 15, 39]. Specifically, the difference in marginal distribution across domains can be reduced by explicitly minimizing predefined distance measures, e.g., Bregman Divergence [31], Geodestic Distance [8] and Maximum Mean Discrepancy (MMD) [9]. The most widely used formulation is MMD due to its compactness and solid theoretical foundations. Furthermore, in order to adapt conditional distribution across domains, Long et al. proposed to exploit the true source and pseudo target labels for computing class-wise MMD [25]. This paper utilizes the MMD and class-wise MMD to jointly align the marginal and conditional distributions across domains.

In order to maintain the data properties hidden in original space, transfer discriminative feature learning aims to not only learn the domain-invariant features, but also preserve label consistency, where distances of the embedded representations from the same class are smaller while distances of the embedded representations from different classes are greater, which is essential for the classification task. For example, VDL [12] and VDA [32] were proposed to minimize the intra-class distance of source domain, while JGSA [39], SCA [6] simultaneously minimize the intra-class distance and maximize the inter-class distance of source domain. Furthermore, DICD [15] takes into account those two discriminative information in both domains. This paper only aims to boost the intra-class compactness of both domains for model simplicity, while the performance is not affected since the variance is maximized. As proved by [16], maximizing the inter-class distance is equivalent to the variance maximization.

However, the class discriminative consistency just emphasizes the global relationships of data points, which makes it unable to deal with the multimodally distributed data. In contrast, TSC [22] and TRSC-GJDA [42] proposed to respect the local manifold structure of data based on the instance graph regularization, GTL [24] further constructed the feature graph regularization to preserve the local manifold structure on the feature side. ARTL [23] and MEDA [35] utilized the Representer theorem in the reproducing kernel Hilbert space to exploit local manifold structure in the label space. In order to highlight the contributions in this paper, our goal is only to exploit the local neighbored relationships between the data instances.

Although these methods could respect the LMS or GDC properties during the transfer feature learning process, their integration is under insufficient exploration so far. Therefore, their result domain-invariant features are either short of discriminative ability, or sensitive to the multimodally distributed data. Moreover, a shortcoming shared by LMS-wise DA methods is that the neighbors leveraged in original data space are not reliable to reveal the intrinsic local structure, especially when the data noises are large. In contrast, this paper takes advantages of both the LMS and GDC, and proposes a novel approach, referred to as Adaptive Local Neighbors for Transfer Discriminative Feature Learning (ALN-TDFL), where the data points' neighbors are revealed adaptively in the desired subspace.

3 Proposed Model

3.1 Notations

In this paper, the bold-italic lowercase letter denotes a vector (e.g., $\boldsymbol{x}$) and bold-italic uppercase letter denotes a matrix (e.g., $\boldsymbol{X}$). The samples from c-th class in source or target domains are defined by $\boldsymbol{X}^{(s/t,c)}$. The superscript $\top$ denotes transpose operator, $tr(\bullet)$ denotes matrix trace operator. $\boldsymbol{I}$ denotes identity matrix and $\boldsymbol{1}$ denotes a matrix whose elements are all 1. $\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_{n\times n}$ is the centering matrix. $||\bullet||_F$ and $||\bullet||_2$ denote l_F-norm and l_2-norm. $\boldsymbol{X}_{i\bullet}$ denotes the i-th row and $\boldsymbol{X}_{ij}$ denotes the element from i-th row and j-th column.

Domain: A domain D consists of a feature space Ω and a marginal distribution $P(\boldsymbol{x})$, and can be formulated as $D = \{\Omega, P(\boldsymbol{x})\}, \boldsymbol{x} \in \Omega$.

Task: Given a specific D, a task consists of a label space Υ and a labeling function $f(\boldsymbol{x})$. From the probabilistic viewpoint, $f(\boldsymbol{x})$ can be interpreted as the conditional distribution $Q(\boldsymbol{y}|\boldsymbol{x})$. Thus, a task can be formulated as $T = \{\Upsilon, Q(\boldsymbol{y}|\boldsymbol{x})\}, \boldsymbol{y} \in \Upsilon$.

Feature-wise Domain Adaptation: Given a labeled source domain $D^{(s)}$ but an unlabeled target domain $D^{(t)}$, assuming that the feature space $\Omega^{(s)} = \Omega^{(t)}$ and label space $\Upsilon^{(s)} = \Upsilon^{(t)}$ while the marginal distribution $P^{(s)}(\boldsymbol{x}^{(s)}) \neq P^{(t)}(\boldsymbol{x}^{(t)})$ and conditional distribution $Q^{(s)}(\boldsymbol{y}^{(s)}|\boldsymbol{x}^{(s)}) \neq Q^{(t)}(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)})$, feature-wise domain adaptation aims to find a projection $\boldsymbol{A}$ to map $D^{(s)}, D^{(t)}$ into a shared subspace where their marginal and conditional distribution differences across domains are explicitly reduced.

3.2 Adaptive Local Neighbors for Transfer Discriminative Feature Learning

In this section, we first introduce how to align the cross-domain features as JDA did [25]. Then the transfer discriminative feature learning (TDFL) is elaborated, which aims to exploit the domain-invariant and discriminative features. Finally, the deficiency of TDFL in multimodally distributed data is discussed, and a novel DA approach to address this challenge is proposed. Most importantly, the neighbored relationships between data points are revealed adaptively in the desired subspace, so that the model is insensitive to the data noises.

3.2.1 Cross-Domain Feature Alignment Revisit

First of all, we represent the source and target data as a data matrix respectively, i.e., $\boldsymbol{X}^{(s)} \in \mathbf{R}^{m\times n^{(s)}}, \boldsymbol{X}^{(t)} \in \mathbf{R}^{m\times n^{(t)}}$, where m is the feature dimension and $n^{(s)}, n^{(t)}$ are the number of samples. The whole data matrix is $\boldsymbol{X} = [\boldsymbol{X}^{(s)}, \boldsymbol{X}^{(t)}] \in \mathbf{R}^{m\times n}$, where $n = n^{(s)} + n^{(t)}$. For leveraging a shared subspace between domains, we define a projection $\boldsymbol{A} \in \mathbf{R}^{m\times k}$, then the new data representation $\boldsymbol{Z} = \boldsymbol{A}^\top \boldsymbol{X}, \boldsymbol{Z} \in \mathbf{R}^{k\times n}$, and dimension is $k (k \ll m)$.

This paper conducts the cross-domain feature alignment by jointly reducing the marginal and conditional distributions across domains. The objective function is defined as follows:

$$\min_{\boldsymbol{A}} \boldsymbol{L}_{MMD} + \beta||\boldsymbol{A}||_F^2 \quad s.t. \quad \boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^\top \boldsymbol{A} = \boldsymbol{I}_k, \tag{1}$$

where β is the trade-off parameter, and the constraint $\boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^\top \boldsymbol{A} = \boldsymbol{I}_k$ enables the data on the subspace are statistically uncorrelated, and $||\boldsymbol{A}||_F^2$ controls the scale of $\boldsymbol{A}$. Now we present the definitions of each term in Eq. 1 in detail. Similar to previous work [25], it utilizes the MMD and class-wise MMD to measure distributional distances across domains. Thus, the distributional differences reduction are equivalent to MMD and class-wise MMD terms minimization. Specifically, MMD is utilized to measure the marginal distribution distance across domains (i.e., $\boldsymbol{P}^{(s)}(\boldsymbol{x}^{(s)}), \boldsymbol{P}^{(t)}(\boldsymbol{x}^{(t)})$) and it computes the deviation between the means of their embedded data as follows:

$$||\frac{1}{n^{(s)}}\sum_{i=1}^{n^{(s)}} \boldsymbol{A}^\top \boldsymbol{x}_i - \frac{1}{n^{(t)}}\sum_{j=1}^{n^{(t)}} \boldsymbol{A}^\top \boldsymbol{x}_j||_2^2 = tr(\boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{M}_0\boldsymbol{X}^\top \boldsymbol{A}), \tag{2}$$

where $\boldsymbol{M}_0$ is the MMD matrix, and it is computed as follows:

$$(\boldsymbol{M}_0)_{ij} = \begin{cases} \frac{1}{n^{(s)}n^{(s)}}, (\boldsymbol{x}_i, \boldsymbol{x}_j \in D^{(s)}) \\ \frac{1}{n^{(t)}n^{(t)}}, (\boldsymbol{x}_i, \boldsymbol{x}_j \in D^{(t)}) \\ -\frac{1}{n^{(s)}n^{(t)}}, (otherwise). \end{cases} \tag{3}$$

Furthermore, the class-wise MMD is utilized to compute the conditional distribution distance across domains as follows (i.e., $\boldsymbol{Q}^{(s)}(\boldsymbol{y}^{(s)}|\boldsymbol{x}^{(s)}), \boldsymbol{Q}^{(t)}(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)})$):

$$\begin{aligned} &\sum_{c=1}^{C} ||\frac{1}{n^{(s,c)}}\sum_{i=1}^{n^{(s,c)}} \boldsymbol{A}^\top \boldsymbol{x}_i - \frac{1}{n^{(t,c)}}\sum_{j=1}^{n^{(t,c)}} \boldsymbol{A}^\top \boldsymbol{x}_j||_2^2 \\ &= \sum_{c=1}^{C} tr(\boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{M}_c\boldsymbol{X}^\top \boldsymbol{A}), \end{aligned} \tag{4}$$

where $n^{(s,c)}$ and $n^{(t,c)}$ are the numbers of data samples belonging to class c in the source and target domains (i.e., $D^{(s,c)}$,$D^{(t,c)}$,$c \in 1, ..., C$), and the class-wise $\boldsymbol{M}_c$ is computed as follows:

$$(\boldsymbol{M}_c)_{ij} = \begin{cases} \frac{1}{n^{(s,c)}n^{(s,c)}}, (\boldsymbol{x}_i, \boldsymbol{x}_j \in D^{(s,c)}) \\ \frac{1}{n^{(t,c)}n^{(t,c)}}, (\boldsymbol{x}_i, \boldsymbol{x}_j \in D^{(t,c)}) \\ -\frac{1}{n^{(s,c)}n^{(t,c)}}, \begin{cases} \boldsymbol{x}_i \in D^{(s,c)}, \boldsymbol{x}_j \in D^{(t,c)} \\ \boldsymbol{x}_j \in D^{(s,c)}, \boldsymbol{x}_i \in D^{(t,c)} \end{cases} \\ 0, (otherwise). \end{cases} \tag{5}$$

Then,

$$\boldsymbol{L}_{MMD} = tr(\boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{M}_0\boldsymbol{X}^\top \boldsymbol{A}) + \sum_{c=1}^{C} tr(\boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{M}_c\boldsymbol{X}^\top \boldsymbol{A}). \tag{6}$$

3.2.2 Adaptive Local Neighbors for Transfer Discriminative Feature Learning

To establish an effective loss term to further prompt the discriminative power of the learned domain-invariant features, we expect to strengthen the inter-class dispersion in both domains, where the distances between the same class instances should be smaller. The formulation is as follows:

$$\begin{aligned} \boldsymbol{L}_{same}^{(s/t)} &= \sum_{c=1}^{C} \frac{1}{n^{(s/t,c)}} \sum_{i,j=1}^{n^{(s/t,c)}} ||\boldsymbol{A}^\top \boldsymbol{x}_i - \boldsymbol{A}^\top \boldsymbol{x}_j||_2^2 \\ &= tr(\boldsymbol{A}^\top \boldsymbol{X}^{(s/t)} \boldsymbol{G}^{(s/t)} {\boldsymbol{X}^{(s/t)}}^\top \boldsymbol{A}), \end{aligned} \tag{7}$$

where $\boldsymbol{G}^{(s/t)} \in \mathbf{R}^{n^{(s/t)}\times n^{(s/t)}}$ is a Laplacian matrix and $\boldsymbol{V}^{(s/t)}$ is computed as follows:

$$\boldsymbol{V}_{ij}^{(s/t)} = \begin{cases} \frac{1}{n^{(s/t,c)}}, (\boldsymbol{y}_i^{(s/t)} = \boldsymbol{y}_j^{(s/t)} = c) \\ 0, (otherwise). \end{cases} \tag{8}$$

Define $d_i = \sum_j \boldsymbol{V}_{ij}^{(s/t)}$, $\boldsymbol{D}^{(s/t)} = diag(d_1, ..., d_{n^{(s/t)}})$, thus $\boldsymbol{G}^{(s/t)} = \boldsymbol{D}^{(s/t)} - \boldsymbol{V}^{(s/t)}$. Then the objective function for transfer discriminative feature learning (TDFL) is defined as follows:

$$\begin{aligned} \min_{\boldsymbol{A}} \boldsymbol{L}_{MMD} + \alpha(\boldsymbol{L}_{same}^{(s)} + \boldsymbol{L}_{same}^{(t)}) + \beta||\boldsymbol{A}||_F^2 \\ s.t. \quad \boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^\top \boldsymbol{A} = \boldsymbol{I}_k. \end{aligned} \tag{9}$$

However, it can be clearly seen from Eq. 7 that TDFL just emphasizes the global relationship of data, which makes it unable to discover the local manifold structure, thus fails to deal with the multimodally distributed data. In order to address this drawback, we propose to incorporate the local manifold structure into the process of transfer discriminative feature learning.

Our motivation is to pull the similar points as closer as possible from the same class. The new loss term of Eq. 7 is defined as:

$$\begin{aligned} \boldsymbol{L}_{same}^{(s/t)*} &= \sum_{c=1}^{C} n^{(s/t,c)} \sum_{i,j=1}^{n^{(s/t,c)}} {\boldsymbol{W}_{ij}^{(s/t,c)}}^2 ||\boldsymbol{A}^\top \boldsymbol{x}_i - \boldsymbol{A}^\top \boldsymbol{x}_j||_2^2 \\ &= tr(\boldsymbol{A}^\top \boldsymbol{X}^{(s/t)} \boldsymbol{G}^{(s/t)*} {\boldsymbol{X}^{(s/t)}}^\top \boldsymbol{A}), \end{aligned} \tag{10}$$

where the matrix $\boldsymbol{W}^{(s/t)} \in \mathbf{R}^{n^{(s/t)}\times n^{(s/t)}}$ is introduced to capture the local relationship between data points, and $\boldsymbol{W}_{ij}^{(s/t,c)} = 0$ if $\boldsymbol{y}_i^{(s/t)} \neq \boldsymbol{y}_j^{(s/t)}$. Likewise, $\boldsymbol{G}^{(s/t)*} = \boldsymbol{D}^{(s/t)*} - \boldsymbol{V}^{(s/t)*}$, where $\boldsymbol{V}^{(s/t)*}$ is computed as follows:

$$\boldsymbol{V}_{ij}^{(s/t)*} = \begin{cases} n^{(s/t,c)} {\boldsymbol{W}_{ij}^{(s/t,c)}}^2, (\boldsymbol{y}_i^{(s/t)} = \boldsymbol{y}_j^{(s/t)} = c) \\ 0, (otherwise). \end{cases} \tag{11}$$

3.2.3 Overall Objective Function

Then the proposed ALN-TDFL is defined as follows:

$$\begin{aligned} \min_{\boldsymbol{A}, \boldsymbol{W}^{(s/t)}} \boldsymbol{L}_{MMD} + \alpha(\boldsymbol{L}_{same}^{(s)*} + \boldsymbol{L}_{same}^{(t)*}) + \beta||\boldsymbol{A}||_F^2 \\ s.t. \quad \forall i, \boldsymbol{W}^{(s/t)} \succeq \boldsymbol{0}, \sum_j \boldsymbol{W}_{ij}^{(s/t)} = 1, \boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^\top \boldsymbol{A} = \boldsymbol{I}_k, \end{aligned} \tag{12}$$

where the constraints on $\boldsymbol{W}^{(s/t)}$ avoid the case that some rows are all zeros. It is noteworthy that the TDFL is a special case of ALN-TDFL, since the Eq. 11 would be degenerated to the Eq. 8 if elements $\boldsymbol{W}_{ij}^{(s/t,c)}$ ($\boldsymbol{y}_i^{(s/t)} = \boldsymbol{y}_j^{(s/t)}$) all equal to $1/n^{(s/t,c)}$.

Supposing the projection $\boldsymbol{A}$ is already obtained, then the weight $\boldsymbol{W}_{ij}^{(s/t,c)}$ would be large if the transformed distances between $\boldsymbol{A}\boldsymbol{x}_i$ and $\boldsymbol{A}\boldsymbol{x}_j$ from the class c are small, which means they are more similar than other data points in the learned subspace. In a later step, if we fix the weight matrix and exploit projection matrix $\boldsymbol{A}$ again, the aim is to emphasize the similar points in the previously learned subspace. Therefore, the points' relationships can be revealed in the desired subspace and it is insensitive to the data noises. Moreover, the challenge of multimodally distributed data in TDFL can be mitigated, since we only focus on the class discriminative consistency among the points' local neighbors.

3.3 Optimization Strategy

Here an adaptive learning strategy is presented to solve the Eq. 12. As for the weight matrix $\boldsymbol{W}^{(s/t)}$, the weight of the points in the class c is initialized as $1/n^{(i,c)}$, while the weight of points from different classes is set to 0. Then the optimal solution can be computed by solving $\boldsymbol{A}$, $\boldsymbol{W}^{(s/t)}$ iteratively.

Domain-invariant feature learning:

When $\boldsymbol{W}^{(s/t)}$ is fixed, then Eq. 12 becomes:

$$\begin{aligned}&\min_{\boldsymbol{A}} tr(\boldsymbol{A}^\top(\boldsymbol{X}\sum_{c=0}^{C}\boldsymbol{M}_c\boldsymbol{X}^\top+\alpha\boldsymbol{X}^{(s)}\boldsymbol{G}^{(s)*}{\boldsymbol{X}^{(s)}}^\top+\\&\alpha\boldsymbol{X}^{(t)}\boldsymbol{G}^{(t)*}{\boldsymbol{X}^{(t)}}^\top)\boldsymbol{A})+\beta||\boldsymbol{A}||_F^2\quad s.t.\quad \boldsymbol{A}^\top\boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^\top\boldsymbol{A}=\boldsymbol{I}_k.\end{aligned} \tag{13}$$

The Eq. 13 is equivalent to a generalised eigen-decomposition problem as follows:

$$\begin{aligned}&(\boldsymbol{X}\sum_{c=0}^{C}\boldsymbol{M}_c\boldsymbol{X}^\top+\alpha\boldsymbol{X}^{(s)}\boldsymbol{G}^{(s)*}{\boldsymbol{X}^{(s)}}^\top+\\&\alpha\boldsymbol{X}^{(t)}\boldsymbol{G}^{(t)*}{\boldsymbol{X}^{(t)}}^\top)+\beta\boldsymbol{I}_m)\boldsymbol{A}=\boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^\top\boldsymbol{A}\Theta,\end{aligned} \tag{14}$$

where $\Theta\in\mathbf{R}^{k\times k}$ is a diagonal matrix with Lagrange Multipliers. The Eq. 14 can be effectively and efficiently solved by calculating the eigenvectors corresponding to the k-smallest eigenvalues.

Adaptive Local Neighbors Learning:

When $\boldsymbol{A}$ and $\boldsymbol{W}^{(t)}$ are fixed, then Eq. 12 becomes:

$$\begin{aligned}&\min_{\boldsymbol{W}^{(s)}}\sum_{c=1}^{C}n^{(s,c)}\sum_{i,j=1}^{n^{(s,c)}}{\boldsymbol{W}_{ij}^{(s,c)}}^2||\boldsymbol{A}^\top\boldsymbol{x}_i-\boldsymbol{A}^\top\boldsymbol{x}_j||_2^2\\&s.t.\quad\forall i,\quad\boldsymbol{W}^{(s)}\succeq\boldsymbol{0},\quad\sum_j\boldsymbol{W}_{ij}^{(s)}=1.\end{aligned} \tag{15}$$

Note that the problem Eq. 15 is independent between different c and i, so we can solve the following problem individually:

$$\begin{aligned}&\min_{\boldsymbol{W}_{i\bullet}^{(s,c)}}\sum_{j=1}^{n^{(s,c)}}{\boldsymbol{W}_{ij}^{(s,c)}}^2||\boldsymbol{A}^\top\boldsymbol{x}_i-\boldsymbol{A}^\top\boldsymbol{x}_j||_2^2\\&s.t.\quad\forall j,\quad\boldsymbol{W}_{ij}^{(s,c)}\geq 0,\quad\sum_j\boldsymbol{W}_{ij}^{(s,c)}=1.\end{aligned} \tag{16}$$

According to [16], the optimal solution to the Eq. 16 is as follows:

$$\boldsymbol{W}_{ij}^{(s,c)}=\frac{1}{||\boldsymbol{A}^\top\boldsymbol{x}_i-\boldsymbol{A}^\top\boldsymbol{x}_j||_2^2}\times\Big(\sum_{t=1}^{n^{(s,c)}}\frac{1}{||\boldsymbol{A}^\top\boldsymbol{x}_i-\boldsymbol{A}^\top\boldsymbol{x}_t||_2^2}\Big)^{-1}. \tag{17}$$

Similarly, the optimal solution to $\boldsymbol{W}_{ij}^{(t,c)}$ is as follows:

$$\boldsymbol{W}_{ij}^{(t,c)}=\frac{1}{||\boldsymbol{A}^\top\boldsymbol{x}_i-\boldsymbol{A}^\top\boldsymbol{x}_j||_2^2}\times\Big(\sum_{t=1}^{n^{(t,c)}}\frac{1}{||\boldsymbol{A}^\top\boldsymbol{x}_i-\boldsymbol{A}^\top\boldsymbol{x}_t||_2^2}\Big)^{-1}. \tag{18}$$

By optimizing $\boldsymbol{A}$ and $\boldsymbol{W}^{(s/t)}$ iteratively, the proposed approach is capable of reducing the distributional differences across domains, prompting the domain-invariant features more discriminative and quantifying the data points' local relationship in the desired subspace. Unlike existing TDFL algorithms, our method integrates the local manifold structure into the TDFL framework, so that it is robust to the multimodally distributed data and insensitive to the data noises. In addition, the proposed objective could monotonically decrease in each iteration, and converge to the lower bound accordingly. A complete procedure of ALN-TDFL is summarized in Algorithm 1.

Algorithm 1: ALN-TDFL

Input: Source data $\boldsymbol{X}^{(s)}$, target data $\boldsymbol{X}^{(t)}$, source labels $\boldsymbol{Y}^{(s)}$, subspace dimensions k, regularized parameters α, β, iterations T
Output: Target labels $\boldsymbol{Y}^{(t)}$
Begin
Initialization
Line 1: Predict $\boldsymbol{Y}^{(t)}$ by some base classifier
Line 2: Compute $\boldsymbol{M}_0$ by Eq. 3
Line 3: Initialize $\boldsymbol{W}^{(s/t)}$ and compute $\boldsymbol{G}^{(s/t)*}$ by Eq. 10
For t=1 to T **do**
Line 4: Update $\sum_{c=1}^{C}\boldsymbol{M}_c$ by Eq. 5
Line 5: Update the projection $\boldsymbol{A}$ by Eq. 14 and the transferred features $\boldsymbol{Z}^{(s)}=\boldsymbol{A}^\top\boldsymbol{X}^{(s)}$, $\boldsymbol{Z}^{(t)}=\boldsymbol{A}^\top\boldsymbol{X}^{(t)}$
Line 6: Update $\boldsymbol{W}^{(s/t)}$ by Eq. 17 and 18
Line 7: Update $\boldsymbol{G}^{(s/t)*}$ by Eq. 10
Line 8: The classifier trained on $\boldsymbol{Z}^{(s)}$, then predict $\boldsymbol{Z}^{(t)}$ and update $\boldsymbol{Y}^{(t)}$
End repeat
Return Target labels $\boldsymbol{Y}^{(t)}$

3.4 Computational Complexity

We analyze the computational complexity of Algorithm 1 using the O notation. We denote T as the number of iterations. The computational cost is detailed as follows: $O(TCn^2)$ for constructing the $\boldsymbol{M}_c$ matrix, i.e., Line 4; $O(Tkm^2)$ for solving the generalized eigen-decomposition problem, i.e., Line 5; $O(T(\sum_c {n^{(s,c)}}^2+\sum_c {n^{(t,c)}}^2))$ for updating $\boldsymbol{W}^{(s/t)}$ i.e., Line 6; In summary, the overall computational complexity of Algorithm 1 is $O(TCn^2+Tkm^2+T(\sum_c {n^{(s,c)}}^2+\sum_c {n^{(t,c)}}^2))$. Moreover, the value of k is not greater than 200, and T is not larger than 20, thus $k,T\ll min(m,n)$. Consequently, Algorithm 1 can be solved in polynomial time concerning the number of samples n.

4 Experiments

4.1 Datasets and Experimental Settings

We adopted the benchmark datasets Office10+Caltech10, Office-Home, Office31, and Image-CLEF-DA in cross-domain image classification to validate the effectiveness of the proposed approach. Fig. 2 and Fig. 3 illustrate some sample images from Office10+Caltech10 and Office-Home datasets, and they follow very different distributions. The dataset descriptions are introduced as follows:

Office10+Caltech10: It contains 2533 images from 10 categories, that forms 4 domains: (A) Amazon, (D) Dslr, (W) Webcam, and (C) Caltech. Then $4\times 3=12$ DA tasks could be constructed, namely $A\rightarrow W$, $C\rightarrow D$ and so on. Note that the arrow "$\rightarrow$" in this paper is the direction from source to target. For instance, $W\rightarrow D$ means Webcam is the labeled source while Dslr is the unlabeled target. Moreover, the Surf features with 800 dimensions [8] and deep feature with 4096 dimensions are adopted, where the deep features

Table 1. Accuracy (%) on the Office10+Caltech10 dataset with Surf and Fc6 features

	Surf features												
Source	C			A			W			D			
Target	A	W	D	C	W	D	C	A	D	C	A	W	Avg.
JDA [25]	43.1	39.3	49.0	40.9	38.0	42.0	33.0	29.8	92.4	31.2	33.4	89.2	46.8
ARTL [23]	44.1	31.5	39.5	36.1	33.6	36.9	29.7	38.3	87.9	30.5	34.9	88.5	44.3
MEDA [35]	56.5	53.9	50.3	43.9	53.2	45.9	34.0	**42.7**	88.5	34.9	41.2	87.5	52.7
VDL [12]	51.0	42.2	45.1	41.5	40.0	38.9	34.2	38.6	82.6	36.4	38.4	82.4	47.6
VDA [32]	46.1	46.1	51.6	42.2	51.2	48.4	27.6	26.1	89.2	31.3	37.7	90.9	49.0
SCA [6]	45.6	40.0	47.1	39.7	34.9	39.5	31.1	30.0	87.3	30.7	31.6	84.4	45.2
JGSA [39]	51.5	45.4	45.9	41.5	45.8	**47.1**	33.2	39.9	90.5	29.9	38.0	91.9	50.1
DICD [15]	47.3	46.4	49.7	42.4	45.1	38.9	33.6	34.1	89.8	34.6	34.5	91.2	49.0
ALN-TDFL	**59.7**	**57.3**	**56.1**	**47.6**	**54.6**	**47.1**	**35.9**	40.8	**94.3**	**37.4**	**42.8**	**95.3**	**55.7**
	Fc6 features												
JDA [25]	89.6	85.1	89.8	83.6	78.3	80.3	84.8	90.3	**100.0**	85.5	91.7	99.7	88.2
DMM [2]	92.4	87.5	90.4	84.8	84.7	**92.4**	81.7	86.5	98.7	83.3	90.7	99.3	89.4
ARTL [23]	92.4	87.8	86.6	87.4	88.5	85.4	88.2	92.3	**100.0**	87.3	92.7	**100.0**	90.7
MEDA [35]	**93.4**	**95.6**	91.1	87.4	88.1	88.1	**93.2**	**99.4**	99.4	87.5	93.2	97.6	92.8
VDA [32]	92.2	82.7	87.3	86.2	80.7	81.5	87.8	91.8	**100.0**	**88.6**	92.9	99.7	89.3
SCA [6]	89.5	85.4	87.9	78.8	75.9	85.4	74.8	86.1	**100.0**	78.1	90.0	98.6	85.9
JGSA [39]	91.4	86.8	93.6	84.9	81.0	88.5	85.0	90.7	**100.0**	86.2	92.0	99.7	90.0
DICD [15]	91.0	92.2	93.6	86.2	81.4	83.4	87.0	89.7	**100.0**	86.1	92.2	99.0	90.1
ALN-TDFL	93.2	93.9	**96.2**	**88.7**	**92.2**	90.5	88.3	92.4	**100.0**	88.3	**93.5**	99.3	**93.0**

Table 2. Accuracy (%) on the Office-Home dataset with ResNet-50 features

	ResNet-50 features												
Source	Ar			Cl			Pr			Rw			
Target	Cl	Pr	Rw	Ar	Pr	Rw	Ar	Cl	Rw	Ar	Cl	Pr	Avg.
JAN [26]	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 [21]	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
MDD [41]	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	**72.5**	**60.2**	82.3	68.1
TADA [37]	53.1	72.3	77.2	59.1	71.2	**72.1**	59.7	53.1	78.4	72.4	60.0	**82.9**	67.6
BSP [3]	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
TAT [18]	51.6	69.5	75.4	59.4	69.5	68.6	59.5	50.5	76.8	70.9	56.6	81.6	65.8
ALN-TDFL	**57.1**	**76.8**	**78.1**	**61.7**	**72.1**	71.9	**62.3**	**54.5**	**78.9**	70.2	59.1	82.7	**68.8**

are pre-extracted from the AlexNet model (Fc6) [13], and pre-trained on ImageNet.

Office-Home: It was released recently as a more challenging DA dataset [33], crawled through several search engines and online image directories. It consists of 4 different domains: (Ar) Artistic images, (Cl) Clipart images, (Pr) Product images, (Rw) Real-World images. Totally, there have 65 object categories for each domain and 15,500 images in the whole dataset. Likewise, $4 \times 3 = 12$ DA tasks can be constructed, and we adopted the deep features with 2048 dimensions, which pre-extracted from the ResNet-50 model [10] and pre-trained on ImageNet.

Office31: It has 3 domains, namely (A) Amazon, (D) Dslr, (W) Webcam, and contains 4,652 images from 31 categories. Similarly, $3 \times 2 = 6$ DA tasks could be constructed, and we also utilize the ResNet-50 deep features.

ImageCLEF-DA: It has 1800 images organized by selecting the 12 common classes shared by 3 public domains: Caltech-256(C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P), where 6 DA tasks can be created and ResNet-50 deep features are adopted.

4.2 Compared Methods

We compared the proposed approach ALN-TDFL with 17 state-of-the-art DA methods for cross-domain image classification problems, including 9 shallow DA methods (i.e., JDA [25], DMM [2], ARTL [23], MEDA [35], VDL [12], VDA [32], JGSA [39], SCA [6], DICD [15]) and 8 deep DA methods (i.e., JAN [26], CDAN [21], MDD [41], TADA [37], BSP [3], TAT [18], CAN [40], MADA [29]).

The ALN-TDFL involves 4 parameters: α, β, k and T. In the next sections, we provide empirical analysis on parameter sensitivity of α, β and k, which verifies that ALN-TDFL can achieve stable performance under a wide range of parameter values. Then we check the convergence of ALN-TDFL w.r.t., T. In the comparative study, we set $k = 20$, $T = 10$, $\alpha = 0.1$, $\beta = 0.05$ for Office10+Caltech10 and ImageCLEF-DA datasets, and $k = 100$, $T = 10$, $\alpha = 0.5$, $\beta = 0.1$ for Office31 and Office-Home datasets since more categories are involved.

Table 3. Accuracy (%) on the Office31 and ImageCLEF-DA datasets with ResNet-50 features

Source	ResNet-50 features												
	A		D		W		C		I		P		
Target	D	W	A	W	A	D	I	P	C	P	C	I	Avg.
JAN [26]	84.7	85.4	68.6	97.4	70.0	99.8	89.5	74.2	94.7	76.8	91.7	88.0	85.1
CDAN [21]	89.8	**93.1**	70.1	98.2	68.0	**100.0**	90.5	74.5	**97.0**	76.7	93.5	90.6	86.8
CAN [40]	85.5	81.5	65.9	98.2	63.4	99.7	89.5	75.8	94.2	78.2	89.2	87.5	84.1
MADA [29]	87.8	90.1	70.3	97.4	66.4	99.6	88.8	75.2	96.0	75.0	92.2	87.9	85.6
ALN-TDFL	**90.6**	88.3	**74.5**	**98.6**	**74.3**	99.8	**92.0**	**78.0**	95.2	**79.5**	**95.3**	**91.2**	**88.1**

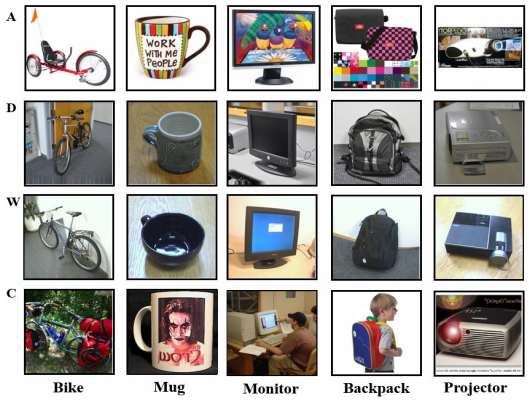

Figure 2. Examplary images from (A) Amazon, (D) Dslr, (W) Webcam and (C) Caltech datasets

Figure 3. Examplary images from (Ar) Art, (Cl) Clipart, (Pr) Product and (Rw) Real-World datasets

4.3 Experimental Results

Comparing with the shallow DA methods, the results on the Office10+Caltech10 with Surf and Fc6 features are shown in Table. 1. It can be seen that our approach outperforms state-of-the-art methods on the most of 24 evaluations. The average classification accuracies of our method are 55.7% and 93.0%, which have 3.0% and 0.2% improvements compared with the best baseline MEDA.

JDA and DMM aim to exploit domain-invariant features by matching the source and target distributions. However, the data structures hidden in the original space are ignored. Therefore, ARTL, MEDA were proposed to respect the local manifold structure (LMS), so that the similar points in original space are closer in the shared subspace.

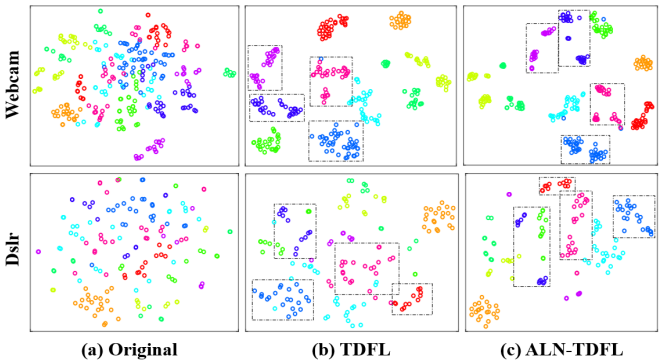

Figure 4. tSNE feature visualization of the task W→D with Surf features. Different classes are denoted by different colors.

Moreover, VDL, VDA, SCA, JGSA, DICD were proposed to preserve global discriminative consistency (GDC) in source or target, so that the domain-invariant features more discriminative.

However, the LMS-wise DA methods are short of discriminative ability, while the GDC-wise DA methods are sensitive to the multimodally distributed data. In the next section, we will explore how GDC damages the shared projection and local data structure. In contrast, we propose to integrate those two metrics for transfer feature learning, thus ALN-TDFL could achieve best results among them.

To further evaluate the effectiveness of the proposed approach, we also report the results of 11 end-to-end deep models. From the results in Tables. 2 and 3, it can be seen that ALN-TDFL also outperforms the deep DA models. Specifically, our approach achieves 0.7% and 1.3% improvements against the best baselines MDD and CDAN. Compared with traditional shallow DA methods, the deep DA methods usually incorporate feature extraction and knowledge transfer procedures into a mainstream deep network, thus promising results could be obtained. Although the deep network have been proven effective and robust in domain adaptation, some representative approaches show that it is nontrivial to be implemented with deep structure. For instance, the class-wise MMD could be easily realized by matrix operations but it is very tricky in deep networks [15]. Additionally, the results on Office-Home evaluate that the proposed approach could be applied to the large-scale dataset, and favorable accuracy could be achieved accordingly. Furthermore, the proposed approach belongs to shallow DA method, which generally runs faster than deep ones since off-the-shelf features are adopted.

4.4 Ablation Study

In this section, we first verify that ALN-TDFL is insensitive to the multimodally distributed data. Then we explore its robustness to the

Table 4. Accuracy (%) on the corrupted Office10+Caltech10 dataset with Surf features

	Corrupted features													
	100		200		300		400		500		600		700	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
C→A	**57.2**	57.1	55.7	**56.9**	51.8	**52.8**	48.2	**48.5**	40.9	**41.9**	**39.0**	38.7	28.8	**30.2**
C→W	52.2	**53.2**	**55.6**	54.9	45.4	**49.8**	44.8	**46.8**	36.3	**41.7**	**31.5**	29.5	27.1	**28.1**
C→D	**51.6**	**51.6**	49.0	**51.6**	47.1	**49.7**	**48.4**	47.1	39.5	**40.1**	34.4	**37.6**	**22.9**	**22.9**
A→C	43.8	**44.7**	43.4	**44.1**	**41.1**	**41.1**	**39.0**	**39.0**	34.7	**35.0**	**30.7**	30.5	**24.1**	23.3
A→W	46.1	**52.9**	46.8	**53.6**	44.8	**46.4**	32.5	**38.0**	34.6	**34.9**	**29.2**	28.8	**24.8**	23.7
A→D	42.0	**45.2**	48.4	**49.0**	45.2	**47.8**	36.9	**37.6**	**37.6**	**37.6**	26.8	**27.4**	22.3	**22.9**
W→C	**34.1**	33.6	32.1	**33.9**	**32.6**	32.4	**32.1**	31.6	**28.1**	27.9	**24.4**	23.9	20.8	**20.9**
W→A	38.5	**40.0**	**41.1**	40.7	37.2	**38.2**	33.8	**34.9**	**33.7**	33.6	30.7	**31.2**	**26.6**	26.0
W→D	93.0	**93.6**	89.8	**90.5**	**82.2**	81.5	**74.5**	73.9	**70.7**	**70.7**	51.0	**54.1**	37.6	**40.8**
D→C	37.2	**38.2**	37.4	**38.4**	32.9	**35.1**	31.1	**32.6**	28.1	**28.5**	**23.9**	**23.9**	20.3	**20.5**
D→A	39.1	**40.6**	40.3	**42.4**	34.2	**34.9**	31.0	**31.5**	**30.0**	27.7	22.3	**22.7**	**23.6**	23.5
D→W	90.5	**92.5**	85.1	**86.8**	84.1	**85.1**	74.2	**76.6**	64.4	**69.2**	**48.1**	47.8	**31.2**	30.2
Avg.	52.1	**53.6**	52.1	**53.6**	48.2	**49.6**	43.9	**44.8**	39.9	**40.7**	32.7	**33.0**	25.8	**26.1**

data noises. Finally, we inspect the distributional and intra-class distances of the leveraged features.

4.4.1 Robust to Multimodally Distributed Data

Following the work in [20], we visualize the features learned by different methods on the task W→ D with Surf features. The results of feature visualization for Original, TDFL, and ALN-TDFL are illustrated in Fig. 4. Comparing with the original features, TDFL behaves better on the discriminative ability, while the local manifold structure is neglected. As can be seen from Fig. 4(c), those subgroups from the same class (inside the black dashed boxes) are distinguished better, and assembled more tightly than TDFL, so that the local structure could be respected perfectly. In addition, the strict constraint of GDC is relaxed since the distances between those subgroups from the same class are not required to be closer, so that the LMS could be respected and the feature transferability unaffected.

4.4.2 Robust to Data Noises

In order to verify ALN-TDFL is robust to the data noises, we randomly corrupt the features on the dataset of Office10-Caltech10 with Surf features. Specifically, we construct n d-dimensions stochastic vectors ($d \in [100, 700]$) whose elements are generated in the range of 0 and 20 in random. Then we utilize them to randomly replace the features of each sample in both domains, and the random selection is repeated 10 times and average results are adopted.

As shown in Table. 4, "No" represents TDFL which does not consider the local manifold structure and no adaptive weight mechanism, while "Yes" is the ALN-TDFL. It can be seen that the classification accuracies are gradually reduced with the data feature noises increase. However, the ALN-TDFL outperforms the TDFL on most evaluations. Therefore, the proposed approach not only respects the local manifold structure during transfer discriminative feature learning, but also insensitive to data noises.

4.4.3 Distributional and Intra-Class Distances

We further explore how the global discriminative consistency influences the equality of learned projection. In this regard, the distributional and intra-class distances are estimated on the dataset of Office10+Caltech10 with Surf features, and they are evaluated on TDFL ("No") and ALN-TDFL ("Yes"), respectively. As can be seen from Table. 5, the distributional distance of TDFL is greater than ALN-TDFL, since the GDC damages the learned projection so that the domain-invariant features perform badly. Moreover, the intra-class distance of ALN-TDFL is smaller that TDFL, since only the local neighbors are taken into consideration.

Table 5. Distributional Distance (DD) and Intra-Class Distance (ICD) on the dataset of Office10+Caltech10 dataset with Surf features

	C→A		C→W		C→D	
	No	Yes	No	Yes	No	Yes
DD	1.0412	**1.0237**	1.3988	**1.3869**	0.6964	**0.6849**
ICD	21.6163	**18.4867**	17.9555	**16.0707**	17.0910	**15.4356**
	A→C		A→W		A→D	
DD	1.6695	**1.6108**	1.4070	**1.3340**	0.8817	**0.8459**
ICD	21.4524	**17.1677**	16.4750	**14.8471**	15.5223	**14.1853**
	W→C		W→A		W→D	
DD	1.7077	**1.5071**	1.9153	**1.6311**	0.1565	**0.1261**
ICD	17.7229	**11.9652**	16.3777	**11.2977**	9.7296	**8.7915**
	D→C		D→A		D→W	
DD	0.9379	**0.7752**	1.0756	**0.9589**	0.1490	**0.1152**
ICD	16.8781	**10.8613**	15.4668	**9.8988**	9.7208	**8.7271**

4.5 Parameters Sensitivity and Convergence

As for the parameters k, β and α, we only report the results on A→C, W→A and D→W with Surf features, while similar trends on all other cross-domain tasks are not shown due to space limitations. We run ALN-TDFL with varying values of one parameter after fixing the others (i.e., $k = 20$, $T = 10$, $\beta = 0.05$, $\alpha = 0.1$). We plot classification results w.r.t., their different values in Fig. 5(a), (c), (d), and choose $k \in [20, 200]$, $\beta \in [0.01, 0.1]$, $\alpha \in [0.1, 1.0]$. We notice that the good performance of classification has a large range.

We also empirically check the convergence property of ALN-TDFL. Fig. 5(b) shows that the objective values decrease steadily with more iterations and converge within only 10 iterations.

5 Conclusion

In this paper, we introduce a novel DA approach, called the Adaptive Local Neighbors for Transfer Discriminative Feature Learning, which not only leverages discriminative domain-invariant features, but also addresses the challenge of multimodally distributed data by

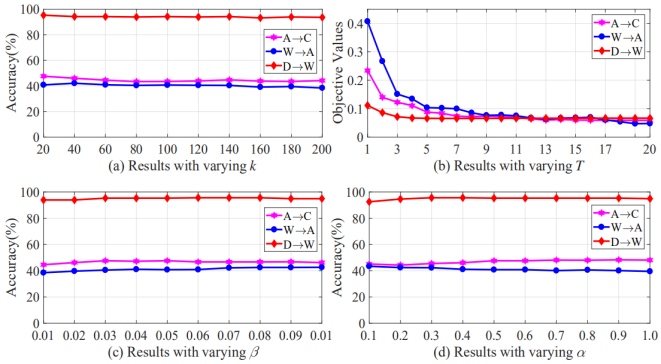

Figure 5. Parameter sensitivity, w.r.t., k, β, α and Convergence w.r.t., T

respecting the local manifold structure. Furthermore, the local neighbors are revealed adaptively that is insensitive to data noises. Extensive experiments show that the proposed approach not only significantly outperforms several state-of-art DA methods, but also obtains desirable results when the data noises exist.

6 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants No.61772108, No.61932020, No.61976038, No.U1908210 and No.61976042.

REFERENCES

[1] S. Bak, P. Carr, and J.F. Lalonde, 'Domain adaptation through synthesis for unsupervised person re-identification', in *ECCV*, pp. 193–209, (2018).
[2] Y. Cao, M.S. Long, and J.M. Wang, 'Unsupervised domain adaptation with distribution matching machines', in *AAAI*, pp. 2795–2802, (2018).
[3] X.Y. Chen, S.N. Wang, M.S. Long, and J.M. Wang, 'Transferability vs. discriminability: batch spectral penalization for adversarial domain adaptation', in *ICML*, pp. 1081–1090, (2019).
[4] W.J. Deng, L. Zheng, Q.X. Ye, G.L. Kang, Y. Yang, and J.B. Jiao, 'Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification', in *CVPR*, pp. 994–1003, (2018).
[5] Z.M. Ding, S. Li, M. Shao, and Y. Fu, 'Graph adaptive knowledge transfer for unsupervised domain adaptation', in *ECCV*, pp. 36–52, (2018).
[6] M. Ghifary, D. Balduzzi, W.B. Kleijn, and M.J. Zhang, 'Scatter component analysis: A unified framework for domain adaptation and domain generalization', *IEEE Trans. Pattern Anal. Mach. Intell.*, **39**(7), 1414–1430, (2017).
[7] B. Gholami, O. Rudovic, and V. Pavlovic, 'Punda: Probabilistic unsupervised domain adaptation for knowledge transfer across visual categories', in *ICCV*, pp. 3601–3610, (2017).
[8] B.Q. Gong, Y. Shi, F. Sha, and K. Grauman, 'Geodesic flow kernel for unsupervised domain adaptation', in *CVPR*, pp. 2066–2073, (2012).
[9] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A.J. Smola, 'A kernel method for the two-sample-problem', in *NeurIPS*, pp. 513–520, (2006).
[10] K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun, 'Deep residual learning for image recognition', in *CVPR*, pp. 770–778, (2016).
[11] L.Q. Hu, M.N. Kan, S.G. Shan, and X.L. Chen, 'Duplex generative adversarial network for unsupervised domain adaptation', in *CVPR*, pp. 1498–1507, (2018).
[12] M. Jiang, W.Z. Huang, Z.Q. Huang, and G.G. Yen, 'Integration of global and local metrics for domain adaptation learning via dimensionality reduction', *IEEE Trans. Cybernetics*, **41**(1), 38–51, (2017).
[13] A. Krizhevsky, I. Sutskever, and G.E. Hinton, 'Imagenet classification with deep convolutional neural networks', in *NeurIPS*, pp. 1106–1114, (2012).
[14] S. Li, C.H. Liu, B.H. Xie, L.M. Su, Z.M. Ding, and G. Huang, 'Joint adversarial domain adaptation', in *ACM MM*, pp. 729–737, (2019).
[15] S. Li, S.J. Song, G. Huang, Z.M. Ding, and C. Wu, 'Domain invariant and class discriminative feature learning for visual domain adaptation', *IEEE Trans. Image Processing*, **27**(9), 4260–4273, (2018).
[16] X.L. Li, M.L. Chen, F.P. Nie, and Q Wang, 'Locality adaptive discriminant analysis', in *IJCAI*, pp. 2201–2207, (2017).
[17] Y. Li, L. Yuan, and N. Vasconcelos, 'Bidirectional learning for domain adaptation of semantic segmentation', in *CVPR*, pp. 6936–6945, (2019).
[18] H. Liu, M.S. Long, J.M. Wang, and M.I. Jordan, 'Transferable adversarial training: A general approach to adapting deep classifiers', in *ICML*, pp. 4013–4022, (2019).
[19] H.F. Liu, M. Shao, Z.M. Ding, and Y. Fu, 'Structure-preserved unsupervised domain adaptation', *IEEE Trans. Knowl. Data Eng.*, **31**(4), 799–812, (2019).
[20] M.S. Long, Y. Cao, J.M. Wang, and M.I. Jordan, 'Learning transferable features with deep adaptation networks', in *ICML*, pp. 97–105, (2015).
[21] M.S. Long, Z.J. Cao, J.M. Wang, and M.I. Jordan, 'Conditional adversarial domain adaptation', in *NeurIPS*, pp. 1647–1657, (2018).
[22] M.S. Long, G.G. Ding, J.M. Wang, J.G. Sun, Y.C. Guo, and P.S. Yu, 'Transfer sparse coding for robust image representation', in *CVPR*, pp. 407–414, (2013).
[23] M.S. Long, J.M. Wang, G.G. Ding, S.J. Pan, and P.S. Yu, 'Adaptation regularization: A general framework for transfer learning', *IEEE Trans. Knowl. Data Eng.*, **26**(5), 1076–1089, (2014).
[24] M.S. Long, J.M. Wang, G.G. Ding, D. Shen, and Q. Yang, 'Transfer learning with graph co-regularization', *IEEE Trans. Knowl. Data Eng.*, **26**(7), 1805–1818, (2014).
[25] M.S. Long, J.M. Wang, G.G. Ding, J.G. Sun, and P.S. Yu, 'Transfer feature learning with joint distribution adaptation', in *ICCV*, pp. 2200–2207, (2013).
[26] M.S. Long, H. Zhu, J.M. Wang, and M.I. Jordan, 'Deep transfer learning with joint adaptation networks', in *ICML*, pp. 2208–2217, (2017).
[27] F.P. Nie, X.Q. Wang, and H. Huang, 'Clustering and projected clustering with adaptive neighbors', in *KDD*, pp. 977–986, (2014).
[28] S.J. Pan and Q. Yang, 'A survey on transfer learning', *IEEE Trans. Knowl. Data Eng.*, **22**(10), 1345–1359, (2010).
[29] Z.Y. Pei, Z.J. Cao, M.S. Long, and J.M. Wang, 'Multi-adversarial domain adaptation', in *AAAI*, pp. 3934–3941, (2018).
[30] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, 'Maximum classifier discrepancy for unsupervised domain adaptation', in *CVPR*, pp. 3723–3732, (2018).
[31] S. Si, D.C. Tao, and B. Geng, 'Bregman divergence-based regularization for transfer subspace learning', *IEEE Trans. Knowl. Data Eng.*, **22**(7), 929–942, (2010).
[32] J. Tahmoresnezhad and S. Hashemi, 'Visual domain adaptation via transfer feature learning', *Knowl. Inf. Syst.*, **50**(2), 585–605, (2017).
[33] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, 'Deep hashing network for unsupervised domain adaptation', in *CVPR*, pp. 5385–5394, (2017).
[34] T.H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, 'Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation', in *CVPR*, pp. 2517–2526, (2019).
[35] J.D. Wang, W.J. Feng, Y.Q. Chen, H. Yu, M.Y. Huang, and P.S. Yu, 'Visual domain adaptation with manifold embedded distribution alignment', in *ACM MM*, pp. 402–410, (2018).
[36] Q. Wang, J.Y. Gao, and X.L. Li, 'Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes', *IEEE Trans. Image Processing*, **28**(9), 4376–4386, (2019).
[37] X.M. Wang, L. Li, W.R. Ye, M.S. Long, and J.M. Wang, 'Transferable attention for domain adaptation', in *AAAI*, pp. 5345–5352, (2019).
[38] L. Wei, S. Zhang, W. Gao, and Q. Tian, 'Person transfer GAN to bridge domain gap for person re-identification', in *CVPR*, pp. 79–88, (2018).
[39] J. Zhang, W.Q. Li, and P. Ogunbona, 'Joint geometrical and statistical alignment for visual domain adaptation', in *CVPR*, pp. 5150–5158, (2017).
[40] W.C. Zhang, W.L. Ouyang, W. Li, and D. Xu, 'Collaborative and adversarial network for unsupervised domain adaptation', in *CVPR*, pp. 3801–3809, (2018).
[41] Y.C. Zhang, T.L. Liu, M.S. Long, and M.I. Jordan, 'Bridging theory and algorithm for domain adaptation', in *ICML*, pp. 7404–7413, (2019).
[42] P. Zhao, W. Wang, Y.J. Lu, H.T. Liu, and S. Yao, 'Transfer robust sparse coding based on graph and joint distribution adaption for image representation', *Knowl. Based Syst.*, **147**, 1–11, (2018).