Discriminative and Geometry Aware Unsupervised Domain Adaptation

Lingkun Luo, Liming Chen, *Senior member, IEEE, Shiqiang Hu, Ying Lu and, Xiaofang Wang,*

Abstract

Domain adaptation (DA) aims to generalize a learning model across training and testing data despite the mismatch of their data distributions. In light of a theoretical estimation of upper error bound, we argue in this paper that an effective DA method should 1) search a shared feature subspace where source and target data are not only aligned in terms of distributions as most state of the art DA methods do, but also discriminative in that instances of different classes are well separated; 2) account for the geometric structure of the underlying data manifold when inferring data labels on the target domain. In comparison with a baseline DA method which only cares about data distribution alignment between source and target, we derive three different DA models, namely **CDDA**, **GA-DA**, and **DGA-DA**, to highlight the contribution of Close yet Discriminative DA(CDDA) based on 1), Geometry Aware DA (**GA-DA**) based on 2), and finally Discriminative and Geometry Aware DA (DGA-DA) implementing jointly 1) and 2). Using both synthetic and real data, we show the effectiveness of the proposed approach which consistently outperforms state of the art DA methods over 36 image classification DA tasks through 6 popular benchmarks. We further carry out in-depth analysis of the proposed DA method in quantifying the contribution of each term of our DA model and provide insights into the proposed DA methods in visualizing both real and synthetic data.

Index Terms

Domain adaptation, Transfer Learning, Visual classification, Discriminative learning, Data distribution matching, Data manifold geometric structure alignment.

I. INTRODUCTION

TRADITIONAL machine learning tasks assume that both training and testing data are drawn from a same data distribution[29], [31], [5]. However, in many real-life applications, due to different factors as diverse as sensor difference, lighting changes, viewpoint variations, *etc.*, data from a target domain may have a different data distribution *w.r.t.* the labeled data in a source domain where a predictor can be can not be reliably learned due to the data distribution shift. On the other hand, manually labeling enough target data for the purpose of training an effective predictor can be very expensive, tedious and thus prohibitive.

Domain adaptation (DA) [29], [31], [5] aims to leverage possibly abundant labeled data from a *source* domain to learn an effective predictor for data in a *target* domain despite the data distribution discrepancy between the source and target. While DA can be *semi-supervised* by assuming a certain amount of labeled data is available in the target domain, in this paper we are interested in *unsupervised* DA[32] where we assume that the target domain has no labels.

State of the art DA methods can be categorized into *instance*-based [29], [7], *feature*-based [30], [22], [42], or *classifier*-based. Classifier-based DA is not suitable to unsupervised DA as it aims to fit a classifier trained on the source data to the target data through adaptation of its parameters, and thereby require some labels in the target domain[38] . The instance-based approach generally assumes that 1) the conditional distributions of source and target domain are identical[44], and 2) certain portion of the data in the source domain can be reused[29] for learning in the target domain through re-weighting. Feature-based adaptation relaxes such a strict assumption and only requires that there exists a mapping from the input data space to a latent shared feature representation space. This latent shared feature space captures the information necessary for training classifiers for source and target tasks. In this paper, we propose a *feature*-based adaptation DA method.

A common method to approach feature adaptation is to seek a low-dimensional latent subspace[31], [30] via dimension reduction. State of the art features two main lines of approaches, namely *data geometric structure alignment*-based or *data distribution* centered. Data geometric structure alignment-based approaches, *e.g.*, **LTSL**[35] , **LRSR**[42], seek a subspace where source and target data can be well aligned and interlaced in preserving inherent hidden geometric data structure via low rank constraint and/or sparse representation. Data distribution centered methods aim to search a latent subspace where the discrepancy between the source and target data distributions is minimized, via various distances, *e.g.*, Bregman divergence[36] based distance, Geodesic distance[13] or Maximum Mean Discrepancy (MMD) [14]. The most popular distance is MMD due to its simplicity and solid theoretical foundations.

A cornerstone theoretical result in DA [2], [17] is achieved by Ben-David *et al.*, who estimate an error bound of a learned hypothesis h on a target domain:

K. Luo, S. Qiang are with School of Aeronautics and Astronautics, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, China e-mail: lolinkun@gmail.com, sqhu@sjtu.edu.cn.

K. Luo, L. Chen, Y. Lv and X. Wang are with LIRIS, CNRS UMR 5205, Ecole Centrale de Lyon, 36 avenue Guy de Collongue, Ecully, France e-mail: (liming.chen,ying.lu,xiaofang.wang,)@ec-lyon.fr.

$$e_{\mathcal{T}}(h) \leq e_{\mathcal{S}}(h) + d_{\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}) +$$
$$\min \left\{ \mathcal{E}_{\mathcal{D}_{\mathcal{S}}} \left[|f_{\mathcal{S}}(\mathbf{x}) - f_{\mathcal{T}}(\mathbf{x})| \right], \mathcal{E}_{\mathcal{D}_{\mathcal{T}}} \left[|f_{\mathcal{S}}(\mathbf{x}) - f_{\mathcal{T}}(\mathbf{x})| \right] \right\} \tag{1}$$

Eq.(1) provides insight on the way to improve DA algorithms as it states that the performance of a hypothesis h on a target domain is determined by: 1) the classification error on the source domain $e_{\mathcal{S}}(h)$; 2) data divergence $d_{\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}})$ which measures the $\mathcal{H}$-*divergence*[17] between two distributions($\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}$); 3) the difference in labeling functions across the two domains. In light of this theoretical result, we can see that data distribution centered DA methods only seek to minimize the second term in reducing data distribution discrepancies, whereas data geometric structure alignment-based methods account for the underlying data geometric structure and expect but without theoretical guarantee the alignment of data distributions.
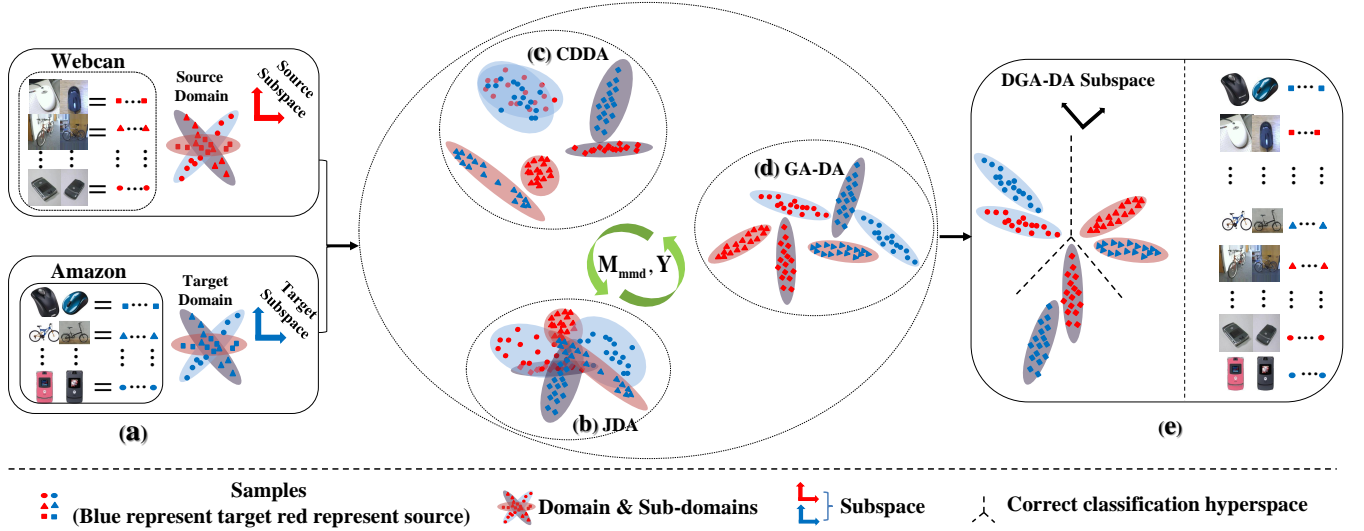

Fig. 1: Illustration of the proposed **DGA-DA** method. Fig.1 (a): source data and target data, *e.g.*, mouse, bike, smartphone images, with different distributions and inherent hidden data geometric structures between the source in red and the target in blue. Samples of different class labels are represented by different geometrical shapes, *e.g.*, round, triangle and square; Fig.1 (b) illustrates **JDA** which closers data distributions whereas **CDDA** (Fig.1 (c)) further makes data discriminative using inter-class repulsive force. Both of them makes use of the nonparametric distance, *i.e.*, Maximum Mean Discrepancy (MMD). Fig.1 (d): accounts for geometric structures of the underlying data manifolds and initial label knowledge in the source domain for label inference; In the proposed DA methods, MMD matrix $\mathbf{M}_{mmd}$ and label matrix $\mathbf{Y}$ are updated iteratively within the processes in Fig.1 (b-d); Fig.1 (e): the achieved latent joint subspace where both marginal and class conditional data distributions are aligned between source and target as well as their data geometric structures; Furthermore, data instances of different classes are well separated from each other, thereby enabling discriminative DA.

In this paper, we argue that an effective DA method should: P1) search a shared feature subspace where source and target data are not only aligned in terms of distributions as most state of the art DA methods do, *e.g.*, **TCA**[28], **JDA**[22], but also *discriminative* in that instances of different classes are well separated; P2) account for the geometric structure of the underlying data manifold when inferring data labels on the target domain.

As a result, we propose in this paper a novel Discriminative Geometry Aware DA (**DGA-DA**) method which provides a unified framework for a simultaneous optimization of the three terms in the upper error bound in Eq.(1). Specifically, the proposed **DGA-DA** also seeks a latent feature subspace to align data distributions as most state of the art DA methods do, but also introduces a *repulsive force* term in the proposed model so as to increase inter-class distances and thereby facilitate discriminative learning and minimize the classification error of the learned hypothesis on source data. Furthermore, the proposed **DGA-DA** also introduces in its model two additional constraints, namely *Label Smoothness Consistency* and *Geometric Structure Consistency*, to account for the geometric structure of the underlying data manifold when inferring data labels in the target domain, thereby minimizing the third term of the error bound of the underlying learned hypothesis on the target domain. Fig.1 illustrates the proposed DA method.

To gain insight into the proposed method and highlight the contribution of P1) and P2) in comparison with a baseline DA method, *i.e.*, **JDA** [22], which only cares about data distribution alignment, we further derive two partial DA methods from our DA model, namely Close yet Discriminative DA (**CDDA**) which implements P1), Geometry Aware DA (**GA-DA**) based on P2), in addition to our Discriminative and Geometry Aware DA (**DGA-DA**) which integrates jointly P1) and P2). Comprehensive experiments carried out on standard DA benchmarks, *i.e.*, 36 cross-domain image classification tasks through 6 datasets, verify the effectiveness of the proposed method, which consistently outperforms the state-of-the-art DA methods. In-depth analysis using both synthetic data and two additional partial models further provide insight into the proposed DA model and highlight its interesting properties.

To sum up, the contributions of this paper are fourfold:

- We propose a novel *repulsive force* term in the DA model to increase the discriminative power of the shared latent subspace, aside from narrowing discrepancies of both the marginal and conditional distributions between the source and target domains.
- We introduce *data geometry awareness*, through Label Smoothness and Geometric Structure Consistencies, for label inference in the proposed DA model and thereby account for the geometric structures of the underlying data manifold.
- We derive from our DA model three novel DA methods, namely **CDDA**, **GA-DA** and **DGA-DA**, which successively implement data discriminativeness, geometry awareness and both, and quantify the contribution of each term beyond a baseline DA method, *i.e.*, **JDA**, which only cares alignment of data distributions.
- We perform extensive experiments on 36 image classification DA tasks through 6 popular DA benchmarks and verify the effectiveness of the proposed method which consistently outperforms twenty-two state-of-the-art DA algorithms with a significant margin. Moreover, we also carry out in-depth analysis of the proposed DA methods, in particular *w.r.t.* their hyper-parameters and convergence speed. In addition, using both synthetic and real data, we also provide insights into the proposed DA model in visualizing the effect of data discriminativeness and geometry awareness.

The paper is organized as follows. Section 2 discusses the related work. Section 3 presents the method. Section 4 benchmarks the proposed DA method and provides in-depth analysis. Section 5 draws conclusion.

II. RELATED WORK

Unsupervised Domain Adaptation assumes no labeled data are provided in the target domain. Thus in order to achieve satisfactory classification performance on the target domain, one needs to learn a classifier with labeled samples provided only from the source domain as well as unlabelled samples from the target domain. In earlier days, this problem is also known as *co-variant shift* and can be solved by sample re-weighting [37]. These methods aim to reduce the distribution difference by re-weighting the source samples according to their relevance to the target samples. While proving useful when the data divergence between the source and target domain is small, these methods fall short to align source and target data when this divergence becomes large.

As a result, recent research in DA has focused its attention on *feature*-based adaptation approach [22], [44], [35], [23], [42], [25], which only assumes a shared latent feature space between the source and target domain. In the learned latent space, the divergence between the projected source and target data distributions is supposed to be minimized. Therefore a classifier learned with the projected labeled source samples could be applied for classification on target samples. To find such a latent shared feature space, many existing methods, *e.g.*,[28], [22], [44], [23], [1], embrace the dimensionality reduction and propose to explicitly minimize some predefined distance measures to reduce the mismatch between source and target in terms of marginal distribution [36] [27] [28], or conditional distribution [33], or both [22]. For example, [36] proposed a Bregman Divergence based regularization schema, which combines Bregman divergence with conventional dimensionality reduction algorithms. In [28], the authors use a similar dimensionality reduction framework while making use of the *Maximum Mean Discrepancy* (MMD) based on the Reproducing Hilbert Space (RKHS) [3] to estimate the distance between distributions. In [22], the authors further improve this work by minimizing not only the mismatch of the cross-domain marginal probability distributions, but also the mismatch of conditional probability distributions.

In line with the focus of manifold learning [45], an increasing number of DA methods, *e.g.*, [24], [35], [42], emphasize the importance of aligning the underlying data manifold structures between the source and the target domain for effective DA. In these methods, low-rank and sparse constraints are introduced into DA to extract a low-dimension feature subspace where target samples can be sparsely reconstructed from source samples [35], or interleaved by source samples [42], thereby aligning the geometric structures of the underlying data manifolds. A few recent DA methods, *e.g.*, **RSA-CDDA**[24], **JGSA**[44], further propose unified frameworks to reduce the shift between domains both statistically and geometrically.

However, in light of the upper error bound as defined in Eq.(1), we can see that data distribution centered DA methods only seek to minimize the second term in reducing data distribution discrepancies, whereas data geometric structure alignment-based methods account for the underlying data geometric structure and expect but without theoretical guarantee the alignment of data distributions. In contrast, the proposed **DGA-DA** method optimizes altogether the three error terms of the upper error bound in Eq.(1).

The proposed **DGA-DA** builds on **JDA** [22] in seeking a latent feature subspace while minimizing the mismatch of both the marginal and conditional probability distributions across domains, thereby decreasing the data divergence term in Eq.(1). But **DGA-DA** goes beyond and differs from **JDA** as we introduce in the proposed DA model a *repulsive force* term so as to increase inter-class distances for discriminative DA, thereby optimizing the first term of the upper error bound in Eq.(1), *i.e.*, the error rate of the learned hypothesis on the source domain. Furthermore, the proposed **DGA-DA** also accounts in its model for the geometric structures of the underlying data manifolds, through label smoothness consistency (LSC) and geometric structure consistency (GSC) which require the inferred labels on the source and target data be smooth and have similar labels on nearby data. These two constraints thus further optimize the third term of the upper error bound in Eq.(1). **DGA-DA** also differs much from a recent DA method, *i.e.*, **SCA**[11], which also tries to introduce data discriminativeness through the between and within class scatter only defined on the source domain. However, besides data geometry awareness that it does not consider, **SCA**

does not seek explicitly data distribution alignment as we do in heritage of **JDA**, nor it has the *repulsive force* term as we introduce in our model in pushing away inter-class data based on both source and target domain. Using both synthetic and real data, sect.IV-F provides insights into and visualizes the differences of the proposed model with a number of state of the art DA methods, *e.g.*, **SCA**, and highlights its interesting properties, in particular data distribution alignment, data discriminativeness and geometry awareness.

III. DISCRIMINATIVE GEOMETRY AWARE DOMAIN ADAPTATION

We first introduce the notations and formalize the problem in sect.III-A, then present in sect.III-B the proposed model for Discriminative and Geometry Aware Domain Adaptation (**DGA-DA**), and solve the model in sect.III-C. Sect.III-D further analyzes the kernelization of the proposed DA model for nonlinear DA problems.

A. Notations and Problem Statement

Matrices are written as boldface uppercase letters. Vectors are written as boldface lowercase letters. For matrix $\mathbf{X} = (x_{ij})$, its i-th row is denoted as $\mathbf{x}^i$, and its j-th column is denoted by $\mathbf{x}_j$. We define the Frobenius norm $\|.\|_F$ as: $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m x_{ij}^2}$.

A domain D is defined as an m-dimensional feature space χ and a marginal probability distribution $P(x)$, *i.e.*, $\mathcal{D} = \{\chi, P(x)\}$ with $x \in \chi$. Given a specific domain D, a task T is composed of a C-cardinality label set $\mathcal{Y}$ and a classifier $f(x)$, *i.e.*, $T = \{\mathcal{Y}, f(x)\}$, where $f(x) = \mathcal{Q}(y|x)$ can be interpreted as the class conditional probability distribution for each input sample x.

In unsupervised domain adaptation, we are given a source domain $\mathcal{D}_\mathcal{S} = \{x_i^s, y_i^s\}_{i=1}^{n_s}$ with n_s labeled samples $\mathbf{X}_\mathcal{S} = [x_1^s...x_{n_s}^s]$, which are associated with their class labels $\mathbf{Y}_S = \{y_1, ..., y_{n_s}\}^T \in \mathbb{R}^{n_s \times c}$, and an unlabeled target domain $\mathcal{D}_\mathcal{T} = \{x_j^t\}_{j=1}^{n_t}$ with n_t unlabeled samples $\mathbf{X}_\mathcal{T} = [x_1^t...x_{n_t}^t]$, whose labels are $\mathbf{Y}_T = \{y_{n_s+1}, ..., y_{n_s+n_t}\}^T \in \mathbb{R}^{n_t \times c}$ are unknown. Here, source domain labels $y_i \in \mathbb{R}^c (1 \le i \le n_s)$ is a binary vector in which $y_i^j = 1$ if x_i belongs to the j-th class. We define the data matrix $\mathbf{X} = [\mathbf{X}_S, \mathbf{X}_T] \in R^{m*n}$ in packing both the source and target data. The source domain $\mathcal{D}_\mathcal{S}$ and target domain $\mathcal{D}_\mathcal{T}$ are assumed to be different, *i.e.*, $\chi_S = \chi_T$, $\mathcal{Y}_S = \mathcal{Y}_T$, $\mathcal{P}(\chi_S) \ne \mathcal{P}(\chi_T)$, $\mathcal{Q}(\mathcal{Y}_S|\chi_S) \ne \mathcal{Q}(\mathcal{Y}_T|\chi_T)$.

We also define the notion of *sub-domain*, denoted as $\mathcal{D}_\mathcal{S}^{(c)}$, representing the set of samples in $\mathcal{D}_\mathcal{S}$ with the label c. Similarly, a sub-domain $\mathcal{D}_\mathcal{T}^{(c)}$ can be defined for the target domain as the set of samples in $\mathcal{D}_\mathcal{T}$ with the label c. However, as samples in the target domain $\mathcal{D}_\mathcal{T}$ are unlabeled, the definition of sub-domains in the target domain, requires a base classifier, *e.g.*, Nearest Neighbor (NN), to attribute pseudo labels for samples in $\mathcal{D}_\mathcal{T}$.

The maximum mean discrepancy (MMD) is an effective non-parametric distance-measure that compares the distributions of two sets of data by mapping the data to Reproducing Kernel Hilbert Space[3] (RKHS). Given two distributions $\mathcal{P}$ and $\mathcal{Q}$, the MMD between $\mathcal{P}$ and $\mathcal{Q}$ is defined as:

$$Dist(P, Q) = \| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(p_i) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(q_i) \|_\mathcal{H} \tag{2}$$

where $P = \{p_1, \ldots, p_{n_1}\}$ and $Q = \{q_1, \ldots, q_{n_2}\}$ are two random variable sets from distributions $\mathcal{P}$ and $\mathcal{Q}$, respectively, and $\mathcal{H}$ is a universal RKHS with the reproducing kernel mapping ϕ: $f(x) = \langle \phi(x), f \rangle$, $\phi : \mathcal{X} \to \mathcal{H}$.

The aim of the Discriminative and Geometry Aware Domain Adaptation (**DGA-DA**) is to learn a latent feature subspace with the following properties: P1) the distances of both marginal and conditional probabilities between the source and target domains are reduced; P2) The distances between each sub-domain to the others are increased so as to increase inter-class distances and thereby enable discriminative DA; and P3) label inference accounts for the underlying data geometric structure.

B. The model

The proposed DA model (sect.III-B5) builds on **TCA** (sect.III-B1) and **JDA** (sect.III-B2) to which discriminative DA (**CDDA**) is introduced (sect.III-B3) and the data geometry awareness (**GA-DA**) is accounted for in label inference and the search of the shared latent feature subspace (sect.III-B4).

1) Search of a Latent Feature Space with Dimensionality Reduction (TCA): The search of a latent feature subspace with dimensionality reduction has been demonstrated useful for DA in several previous works, *e.g.*, [28], [22], [24], [35], [44]. In projecting original raw data into a lower dimensional space, the *principal* data structure is preserved while decreasing its complexities. In the proposed method, we also apply the Principal Component Analysis (PCA) to capture the major data structure. Mathematically, given an input data matrix $\mathbf{X} = [\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}]$, $\mathbf{X} \in \mathbb{R}^{m \times (n_s + n_t)}$, the centering matrix is defined as $\mathbf{H} = \mathbf{I} - \frac{1}{n_s + n_t} \mathbf{1}$, where $\mathbf{1}$ is the $(n_s + n_t) \times (n_s + n_t)$ matrix of ones. The optimization of PCA is to find a projection transformation $\mathbf{A}$ which maximizes the embedded data variance.

$$\max_{\mathbf{A}^T \mathbf{A} = \mathbf{I}} tr(\mathbf{A}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{A}) \tag{3}$$

where $tr(\cdot)$ denotes the trace of a matrix, $\boldsymbol{XHX}^T$ is the data covariance matrix, and $\mathbf{A} \in \mathbb{R}^{\mathbf{m}\times\mathbf{k}}$ with m the feature dimension and k the dimension of the projected subspace. The optimal solution is calculated by solving an eigendecomposition problem: $\boldsymbol{XHX}^T = \boldsymbol{A\Phi}$, where $\boldsymbol{\Phi} = diag(\phi_1, \ldots, \phi_k)$ are the k largest eigenvalues. Finally, the original data $\boldsymbol{X}$ is projected into the optimal k-dimensional subspace using $\boldsymbol{Z} = \boldsymbol{A}^T\boldsymbol{X}$.

2) Joint Marginal and Conditional Distribution Domain Adaptation (JDA): However, the previous feature subspace calculated via PCA does not align explicitly data distributions between the source and target domain. Following [22], [21], we also empirically measure the distance of both marginal and conditional distributions across domain via the nonparametric distance measurement MMD in RKHS [3] once the original data projected into a low-dimensional feature space. Formally, the empirical distance of the two domains is defined as:

$$
Dist^{marginal}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) =
$$
$$
\left\| \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbf{A}^T x_i - \frac{1}{n_t}\sum_{j=n_s+1}^{n_s+n_t} \mathbf{A}^T x_j \right\|^2 = tr(\mathbf{A}^T\mathbf{X}\mathbf{M_0}\mathbf{X}^T\mathbf{A}) \tag{4}
$$

where $\mathbf{M}_0$ represents the marginal distribution between $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ and its calculation is obtained by:

$$
(\mathbf{M}_0)_{ij} = \begin{cases} \frac{1}{n_s n_s}, & x_i, x_j \in D_S \\ \frac{1}{n_t n_t}, & x_i, x_j \in D_T \\ \frac{-1}{n_t n_s}, & otherwise \end{cases} \tag{5}
$$

where $x_i, x_j \in (\mathcal{D}_\mathcal{S} \cup \mathcal{D}_\mathcal{T})$. The difference between the marginal distributions $\mathcal{P}(\mathcal{X}_\mathcal{S})$ and $\mathcal{P}(\mathcal{X}_\mathcal{T})$ is reduced in minimizing $Dist^{marginal}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T})$.

Similarly, the distance of conditional probability distributions is defined as the sum of the empirical distances over the class labels between the sub-domains of a same label in the source and target domain:

$$
Dist^{conditional}\sum_{c=1}^{C}(\mathcal{D}_\mathcal{S}{}^c, \mathcal{D}_\mathcal{T}{}^c) = \left\| \frac{1}{n_s^{(c)}}\sum_{x_i\in\mathcal{D}_\mathcal{S}{}^{(c)}} \mathbf{A}^T x_i - \frac{1}{n_t^{(c)}}\sum_{x_j\in\mathcal{D}_\mathcal{T}{}^{(c)}} \mathbf{A}^T x_j \right\|^2 = tr(\mathbf{A}^T\mathbf{X}\mathbf{M}_c\mathbf{X}^T\mathbf{A}) \tag{6}
$$

where C is the number of classes, $\mathcal{D}_\mathcal{S}{}^{(c)} = \{x_i : x_i \in \mathcal{D}_\mathcal{S} \wedge y(x_i) = c\}$ represents the cth sub-domain in the source domain, $n_s^{(c)} = \left\|\mathcal{D}_\mathcal{S}{}^{(c)}\right\|_0$ is the number of samples in the cth source sub-domain. $\mathcal{D}_\mathcal{T}{}^{(c)}$ and $n_t^{(c)}$ are defined similarly for the target domain. Finally, $\mathbf{M_c}$ represents the conditional distribution between sub-domains in $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ and it is defined as:

$$
(\mathbf{M}_c)_{ij} = \begin{cases} \frac{1}{n_s^{(c)} n_s^{(c)}}, & x_i, x_j \in D_\mathcal{S}{}^{(c)} \\ \frac{1}{n_t^{(c)} n_t^{(c)}}, & x_i, x_j \in D_\mathcal{T}{}^{(c)} \\ \frac{-1}{n_s^{(c)} n_t^{(c)}}, & \begin{cases} x_i \in D_\mathcal{S}{}^{(c)}, x_j \in D_\mathcal{T}{}^{(c)} \\ x_i \in D_\mathcal{T}{}^{(c)}, x_j \in D_\mathcal{S}{}^{(c)} \end{cases} \\ 0, & otherwise \end{cases} \tag{7}
$$

In minimizing $Dist^{conditional}\sum_{c=1}^{C}(D_\mathcal{S}{}^c, D_\mathcal{T}{}^c)$, the mismatch of conditional distributions between $D_\mathcal{S}{}^c$ and $D_\mathcal{T}{}^c$ is reduced.

3) Close yet Discriminative Domain Adaptation (CDDA): However, the previous joint alignment of the marginal and conditional distributions across domain does not explicitly render data discriminative in the searched feature subspace. As a result, we introduce a *Discriminative* domain adaption via a *repulsive force* term, so as to increase the distances of sub-domains with different labels, and improve the discriminative power of the latent shared features, thereby making it possible for a better predictive model for both the source and target data.

Specifically, the *repulsive force* term is defined as: $Dist^{repulsive} = Dist_{\mathcal{S}\rightarrow\mathcal{T}}^{repulsive} + Dist_{\mathcal{T}\rightarrow\mathcal{S}}^{repulsive}$, where $\mathcal{S}\rightarrow\mathcal{T}$ and $\mathcal{T}\rightarrow\mathcal{S}$ index the distances computed from $D_\mathcal{S}$ to $D_\mathcal{T}$ and $D_\mathcal{T}$ to $D_\mathcal{S}$, respectively. $Dist_{\mathcal{S}\rightarrow\mathcal{T}}^{repulsive}$ represents the sum of the distances between each source sub-domain $D_\mathcal{S}{}^{(c)}$ and all the target sub-domains $D_\mathcal{T}{}^{(r); \, r\in\{\{1...C\}-\{c\}\}}$ except the one with the label c. The sum of these distances is explicitly defined as:

$$
Dist_{\mathcal{S}\rightarrow\mathcal{T}}^{repulsive} = \sum_{c=1}^{C} \left\| \frac{1}{n_s^{(c)}}\sum_{x_i\in D_\mathcal{S}{}^{(c)}} \mathbf{A}^T x_i - \frac{1}{\sum_{r\in\{\{1...C\}-\{c\}\}} n_t^{(r)}}\sum_{x_j\in D_\mathcal{T}{}^{(r)}} \mathbf{A}^T x_j \right\|^2 = \sum_{c=1}^{C} tr(\mathbf{A}^T\mathbf{X}\mathbf{M}_{\mathcal{S}\rightarrow\mathcal{T}}\mathbf{X}^T\mathbf{A}) \tag{8}
$$

where $\mathbf{M}_{\mathcal{S}\to\mathcal{T}}$ is defined as

$$(\mathbf{M}_{\mathcal{S}\to\mathcal{T}})_{\mathbf{ij}} = \begin{cases} \frac{1}{n_s^{(c)}n_s^{(c)}}, & x_i, x_j \in D_{\mathcal{S}}{}^{(c)} \\ \frac{1}{n_t^{(r)}n_t^{(r)}}, & x_i, x_j \in D_{\mathcal{T}}{}^{(r)} \\ \frac{-1}{n_s^{(c)}n_t^{(r)}}, & \begin{cases} x_i \in \mathcal{D}_{\mathcal{S}}{}^{(c)}, x_j \in D_{\mathcal{T}}{}^{(r)} \\ x_i \in \mathcal{D}_{\mathcal{T}}{}^{(r)}, x_j \in \mathcal{D}_{\mathcal{S}}{}^{(c)} \end{cases} \\ 0, & otherwise \end{cases} \tag{9}$$

Symmetrically, $Dist_{\mathcal{T}\to\mathcal{S}}^{repulsive}$ represents the sum of the distances from each target sub-domain $D_{\mathcal{T}}{}^{(c)}$ to all the the source sub-domains $D_{\mathcal{S}}{}^{(r);\ r\in\{\{1...C\}-\{c\}\}}$ except the source sub-domain with the label c. Similarly, the sum of these distances is explicitly defined as:

$$Dist_{T\to S}^{repulsive} = \sum_{c=1}^{C} \left\| \frac{1}{n_s^{(c)}} \sum_{x_i \in D_T{}^{(c)}} \mathbf{A}^T x_i - \frac{1}{\sum_{r\in\{\{1...C\}-\{c\}\}} n_t^{(r)}} \sum_{x_j \in D_S{}^{(r)}} \mathbf{A}^T x_j \right\|^2 = \sum_{c=1}^{C} tr(\mathbf{A}^T\mathbf{X}\mathbf{M}_{T\to S}\mathbf{X}^{\mathbf{T}}\mathbf{A}) \tag{10}$$

where $\mathbf{M}_{\mathcal{T}\to\mathcal{S}}$ is defined as

$$(\mathbf{M}_{\mathcal{T}\to\mathcal{S}})_{\mathbf{ij}} = \begin{cases} \frac{1}{n_t^{(c)}n_t^{(c)}}, & x_i, x_j \in D_{\mathcal{T}}{}^{(c)} \\ \frac{1}{n_s^{(r)}n_s^{(r)}}, & x_i, x_j \in D_{\mathcal{S}}{}^{(r)} \\ \frac{-1}{n_t^{(c)}n_s^{(r)}}, & \begin{cases} x_i \in \mathcal{D}_{\mathcal{T}}{}^{(c)}, x_j \in D_{\mathcal{S}}{}^{(r)} \\ x_i \in \mathcal{D}_{\mathcal{S}}{}^{(r)}, x_j \in \mathcal{D}_{\mathcal{T}}{}^{(c)} \end{cases} \\ 0, & otherwise \end{cases} \tag{11}$$

Finally, we obtain

$$Dist^{repulsive} = \sum_{c=1}^{C} tr(\mathbf{A}^T\mathbf{X}(\mathbf{M}_{S\to T} + \mathbf{M}_{T\to S})\mathbf{X}^{\mathbf{T}}\mathbf{A}) \tag{12}$$

We define $\mathbf{M}_{\hat{c}} = \mathbf{M}_{S\to T} + \mathbf{M}_{T\to S}$ as the *repulsive force* matrix. While the minimization of Eq.(4) and Eq.(6) makes closer both marginal and conditional distributions between source and target, the maximization of Eq.(12) increases the distances between source and target sub-domains, thereby improve the discriminative power of the searched latent feature subspace.

4) Geometry Aware Domain Adaptation (GA-DA): In a number of state of the art DA methods, *e.g.*,[27], [28], [22], the simple *Nearest Neighbor* (NN) classifier is applied for label inference. In **JDA** and **LRSR**[42], NN-based label deduction is applied twice at each iteration. NN is first applied to the target domain in order to generate the *pseudo* labels of the target data and enable the computation of the conditional probability distance as defined in sect. III-B2. Once the optimized latent subspace identified, NN is then applied once again at the end of an iteration for the label prediction of the target domain. However, given the neighborhood usually based on the $L2$ or $L1$ distance, NN could fall short to measure the similarity of source and target domain data which may be embedded into a manifold with complex geometric structures.

To account for the underlying data manifold structure in data similarity measurement, we further introduce two consistency constraints, namely *label smoothness consistency* and *geometric structure consistency* for both the *pseudo* and final label inference.

Label Smoothness Consistency (LSC): LSC is a constraint designed to prevent too much changes from the initial query assignment $\mathbf{Y}_{\mathcal{S}}$.

$$Dist^{lable} = \sum_{j=1}^{C} \sum_{i=1}^{n_s+n_t} \left\| \mathbf{Y}_{i,j}^{(F)} - \mathbf{Y}_{i,j}^{(0)} \right\| \tag{13}$$

where $\mathbf{Y} = \mathbf{Y}_{\mathcal{S}} \cup \mathbf{Y}_{\mathcal{T}}$, $\mathbf{Y}_{i,j}^{(F)}$ is the calculated probability of i_{th} data belonging to j_{th} class. Each data x_i has a predicted label $y_i = \arg\max_{j\leq c} \mathbf{Y}_{ij}^F$. $\mathbf{Y}_{i,j}^{(0)}$ is the initial prediction. As for unlabeled target data $\mathbf{X}_{\mathcal{T}}$, traditional ranking methods[18], [43] assign the labels $\mathbf{Y}_T = \mathbf{0}^{n_t*c}$. However, this definition lacks discriminative properties due to the equal probability assignments in $\mathbf{X}_{\mathcal{T}}$. In this work, we define the initial $\mathbf{Y}^{(0)}$ as:

$$\mathbf{Y}_{\mathcal{S}_{(ij)}}^{(0)} = \begin{cases} y_{\mathcal{S}_{(ij)}}^{(0)} = 1 \ (1 \leq i \leq n_s), j = c, y_{ij} \in D_{\mathcal{S}}^{(c)} \\ 0 \qquad else \end{cases}$$

$$\mathbf{Y}_{\mathcal{T}_{(ij)}}^{(0)} = \begin{cases} y_{\mathcal{T}_{(ij)}}^{(0)} = 1 \ ((n_s + 1) \leq i \leq n_s + n_t), j = c, \\ y_{ij} \in D_{\mathcal{T}}^{(c)} \\ 0 \qquad else \end{cases} \tag{14}$$

where $D_{\mathcal{T}}^{(c)}$ is defined as pseudo labels, generated via a base classifier, *e.g.*, NN.

Geometric Structure Consistency (GSC): GSC is designed to ensure that inferred data labels comply with the geometric structures of the underlying data manifolds. We propose to characterize alignment of label inference with the underlying data geometric structure through the Laplace matrix $\mathbf{L}$:

$$\mathbf{Y}^T \mathbf{L} \mathbf{Y} = \mathbf{Y}^T (\mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}) \mathbf{Y} = \sum_{i=1}^{n_s+n_t} d_{ii} \left(\frac{y_i}{\sqrt{\mathbf{d}_{ii}}} \right)^2 - \sum_{i,j=1}^{n_s+n_t} \mathbf{d}_{ii} \left(\frac{y_i}{\sqrt{\mathbf{d}_i}} \frac{y_j}{\sqrt{\mathbf{d}_j}} \right)^2 \mathbf{w}_{ij} = \frac{1}{2} \sum_{i,j=1}^{n_s+n_t} \mathbf{w}_{ij} \left(\frac{y_i}{\sqrt{\mathbf{d}_{ii}}} - \frac{y_j}{\sqrt{\mathbf{d}_{jj}}} \right)^2, \tag{15}$$

where $\mathbf{W} = [w_{ij}]_{(n_s+n_t) \times (n_s+n_t)}$ is an affinity matrix [26], with w_{ij} giving the affinity between two data samples i and j and defined as $w_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ if $i \neq j$ and $w_{ii} = 0$ otherwise, $\mathbf{D} = diag\{d_{11}...d_{(n_s+n_t),(n_s+n_t)}\}$ is the degree matrix with $d_{ii} = \sum_j w_{ij}$. When Eq.(15) is minimized, the geometric structure consistency ensures that the label space does not change too much between nearby data.

5) the final model (DGA-DA): Our final DA model integrates: 1) alignment of both marginal and conditional distributions across domain as defined by Eq.(4) and Eq.(6), 2) the repulsive force as in Eq.(12), and 3) data geometry aware label inference through both the label smoothness (Eq.(13)) and geometric structure (Eq.(15)) consistencies. Therefore, our final model is defined as:

$$\min(Dist^{marginal} + Dist^{conditional} + Dist^{label} + \mathbf{Y}^T L \mathbf{Y}) + \max(Dist^{repulsive}) \tag{16}$$

It can be re-written mathematically as:

$$\min_{\mathbf{A}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{A} = \mathbf{I}} \left(\sum_{c=0}^{C} tr(\mathbf{A}^T \mathbf{X} \mathbf{M}_c \mathbf{X}^T A) + \lambda \|\mathbf{A}\|_F^2 + \mu (\sum_{j=1}^{C} \sum_{i=1}^{n_s+n_t} \left\| \mathbf{Y}_{ij}^{(F)} - \mathbf{Y}_{ij}^{(0)} \right\|) + \mathbf{Y}^T \mathbf{L} \mathbf{Y} \right) + \max_{\mathbf{A}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{A} = \mathbf{I}} tr(\mathbf{A}^T \mathbf{X} \mathbf{M}_{\hat{c}} \mathbf{X}^T \mathbf{A}) \tag{17}$$

where the constraint $\mathbf{A}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{A} = \mathbf{I}$ removes an arbitrary scaling factor in the embedding and prevents the above optimization collapse onto a subspace of dimension less than the required dimensions. λ is a regularization parameter to guarantee the optimization problem to be well-defined. μ is a trade-off parameter which balances LSC and GSC.

C. Solving the model

Direct solution to Eq.(17) is nontrivial. We divide it into two sub-problems.
Sub-problem (a):

$$\min_{\mathbf{A}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{A} = \mathbf{I}, \mathbf{M}_{cyd} = \sum_{c=0}^{C} \mathbf{M}_c - \mathbf{M}_{\hat{c}}} \left(\sum_{c=0}^{C} tr(\mathbf{A}^T \mathbf{X} \mathbf{M}_{cyd} \mathbf{X}^T A) + \lambda \|\mathbf{A}\|_F^2 \right), \tag{18}$$

Sub-problem (b):

$$\min \left(\mu \sum_{j=1}^{C} \sum_{i=1}^{n_s+n_t} \left\| \mathbf{Y}_{ij}^{(F)} - \mathbf{Y}_{ij}^{(0)} \right\| + \mathbf{Y}^T \mathbf{L} \mathbf{Y} \right) \tag{19}$$

These two sub-problems are then iteratively optimized.

Sub-problem (a) amounts to solving the generalized eigendecomposition problem. Augmented Lagrangian method [10], [22] can be used to solve this problem. In setting its partial derivation *w.r.t.* A equal to zero, we obtain:

$$(\mathbf{X} \mathbf{M}_{cyd} \mathbf{X}^T + \lambda \mathbf{I}) \mathbf{A} = \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{A} \Phi \tag{20}$$

where $\Phi = \mathrm{diagram}(\varphi_1, ...\varphi_k) \in R^{k*k}$ is the Lagrange multiplier. The optimal subspace A is reduced to solving Eq.(20) for the k smallest eigenvectors. Then, we obtain the projection matrix A and the underlying embedding space $\mathbf{Z} = \mathbf{A}^T \mathbf{X}$.

Sub-problem (b) is nontrivial. Inspired by the solution proposed in [45] [18] [43], the minimum is approached where the derivative of the function is zero. An approximate solution can be provided by:

$$\mathbf{Y}^{\star} = (\mathbf{D} - (\frac{1}{1+\mu})\mathbf{W})^{-1} \mathbf{Y}^{(0)} \tag{21}$$

where $\mathbf{Y}^{\star}$ is the probability of prediction of the target domain corresponding to different class labels, $\boldsymbol{W}$ is an affinity matrix and $\boldsymbol{D}$ is the diagonal matrix.

For sake of simplicity, we define $\alpha = \frac{1}{1+\mu}$ and then Eq.(21) is reformulated as Eq.(22):

$$\mathbf{Y}^{\star} = (\mathbf{D} - \alpha\mathbf{W})^{-1}\mathbf{Y}^{(0)} \tag{22}$$

To sum up, at a given iteration, sub-problem (a) as in Eq.(18) searches a latent feature subspace $\mathbf{Z}$ in closering both marginal and conditional data distributions between source and target while making use of source and current target labels in pushing away interclass data; sub-problem (b) as in Eq.(19) infers through Eq.(22) novel labels for target data in line with source data labels while making use of the geometric structures of the underlying data manifolds in the current subspace $\mathbf{Z}$. This iterative process eventually ends up in a latent feature subspace where : 1) the discrepancies of both marginal and conditional data distributions between source and target are narrowed; 2) source and target data are rendered more discriminative thanks to the increase of interclass distances; and 3) the geometric structures of the underlying data manifolds are aligned.

The complete learning algorithm is summarized in Algorithm 1 - **DGA-DA**.

Algorithm 1: Discriminative Geometry Aware Domain Adaptation (**DGA-DA**)

Input: Data $\mathbf{X}$, Source domain labels $\mathbf{Y}_{\mathcal{S}}$, subspace dimension k, number of iterations T, regularization parameters λ and α

1 **Step 1**: Initialize the iteration counter t=0 and compute $\mathbf{M}_0$ as in Eq.(5).

2 **Step 2**: Initialize pseudo target labels $\mathbf{Y}_{\mathcal{T}}$ and projection space $\mathbf{A}$:

3 (1) Solve the generalized eigendecomposition problem[10], [22] as in Eq.(20) (replace $\mathbf{M_{cyd}}$ by $\mathbf{M}_0$) and obtain adaptation matrix $\mathbf{A}$, then embed data via the transformation, $\mathbf{Z} = \mathbf{A^T X}$;

4 (2)Initialize pseudo target labels $\mathbf{Y}_{\mathcal{T}}$ via a base classifier, *e.g.*, 1-NN, based on source domain labels $\mathbf{Y}_{\mathcal{S}}$.

5 **while** *not converged and $t < T$* **do**

6   **Step 3**: Update projection space $\mathbf{A}$

7   (i) Compute $\mathbf{M}_c$ (Eq.(7))

8   (ii) Compute $\mathbf{M}_{\hat{c}} = \mathbf{M}_{S \to T} + \mathbf{M}_{T \to S}$ as in Eq.(11) and Eq.(9) via $\mathbf{Y}_{\mathcal{T}}$.

9   (iii) Calculate $\mathbf{M}_{cyd} = \mathbf{M}_c + \mathbf{M}_0 - \mathbf{M}_{\hat{c}}$;

10   (iv) Solve Eq.(20) then update $\mathbf{A}$ and $\mathbf{Z} = \mathbf{A^T X}$;

11   **Step 4**: Label deduction

12   (i) construct the label matrix $\mathbf{Y}^{(0)}$ as in Eq.(14);

13   (ii) design the affinity matrix[26] $\boldsymbol{W}$ and diagonal matrix $\boldsymbol{D}$;

14   (iii) obtain $\mathbf{Y}_{final}$ in solving Eq.(21);

15   **Step 5**: update pseudo target labels $\{\mathbf{Y}_{\mathcal{T}}^{(F)} = \mathbf{Y}_{final}\,[:, (n_s + 1) : (n_s + n_t)]\}$;

16   **Step 6**: $t = t + 1$; Return to Step 3;

Output: Adaptation matrix $\mathbf{A}$, embedding $\mathbf{Z}$, Target domain labels $\mathbf{Y}_{\mathcal{T}}^{(F)}$

D. Kernelization Analysis

The proposed **DGA-DA** method can be extended to nonlinear problems in a Reproducing Kernel Hilbert Space via the kernel mapping $\phi : x \to \phi(x)$, or $\phi(\mathbf{X}) : [\phi(\mathbf{x}_1), ..., \phi(\mathbf{x}_n)]$, and the kernel matrix $\mathbf{K} = \phi(\mathbf{X})^T\phi(\mathbf{X}) \in R^{n*n}$. We utilize the Representer theorem to formulate Kernel **DGA-DA** as

$$\min_{\mathbf{A}^T\mathbf{KHK}^T\mathbf{A}=\mathbf{I}} \left(\sum_{c=0}^{C} tr(\mathbf{A}^T\mathbf{KM}_c\mathbf{K}^T\mathbf{A}) + \lambda\|\mathbf{A}\|_F^2 + \sum_{j=1}^{C}\sum_{i=1}^{n_s+n_t} \left\|\mathbf{Y}_{ij}^{(F)} - \mathbf{Y}_{ij}^{(0)}\right\| + \mathbf{Y}^T\mathbf{LY} \right) + \max_{\mathbf{A}^T\mathbf{KHK}^T\mathbf{A}=\mathbf{I}} tr(\mathbf{A}^T\mathbf{KM}_{\hat{c}}\mathbf{K}^T\mathbf{A}) \tag{23}$$

IV. EXPERIMENTS

In this section, we verify and analyze in-depth the effectiveness of our proposed domain adaptation model, *i.e.*, **DGA-DA**, on 36 cross domain image classification tasks generated by permuting six datasets (see Fig.2). Sect.IV-A describes the benchmarks and the features. Sect.IV-B lists the baseline methods which the proposed **DGA-DA** is compared to. Sect.IV-C presents the experimental setup and introduces in particular two partial DA methods, namely **CDDA** and **GA-DA**, in addition to the proposed **DGA-DA** based on our full DA model. Sect.IV-D discusses the experimental results in comparison with the state of the art. Sect.IV-E analyzes the convergence and parameter sensitivity of the proposed method. Sect.IV-F further provides insight into the proposed DA model in visualizing the achieved feature subspaces through both synthetic and real data.

A. Benchmarks and Features

As illustrated in Fig.2, USPS[15]+MINIST[20], COIL20[22], PIE[22] and office+Caltech[22], [42], [?], [35] are standard benchmarks for the purpose of evaluation and comparison with state-of-the-art in DA. In this paper, we follow the data preparation as most previous works[40], [42], [12], [11], [6], [24] do. We construct 36 datasets for different image classification tasks.

Office+Caltech consists of 2533 images of ten categories (8 to 151 images per category per domain)[11]. These images come from four domains: (A) AMAZON, (D) DSLR, (W) WEBCAM, and (C) CALTECH. AMAZON images were acquired in a controlled environment with studio lighting. DSLR consists of high resolution images captured by a digital SLR camera in a home environment under natural lighting. WEBCAM images were acquired in a similar environment to DSLR, but with a low-resolution webcam. CALTECH images were collected from Google Images.
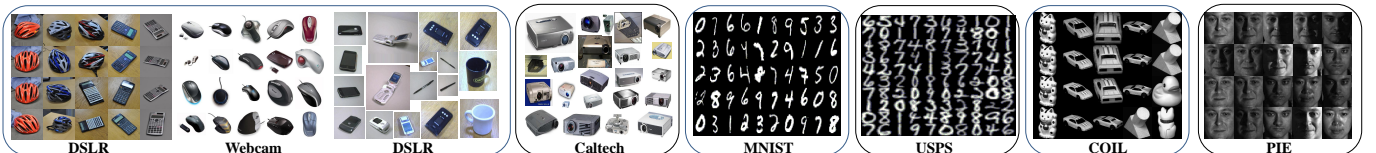
We use two types of image features extracted from these datasets, *i.e.*, **SURF** and **DeCAF6**, that are publicly available. The **SURF**[13] features are *shallow* features extracted and quantized into an 800-bin histogram using a codebook computed with K-means on a subset of images from Amazon. The resultant histograms are further standardized by z-score. The **Deep Convolutional Activation Features (DeCAF6)**[8] are *deep* features computed as in **AELM**[40] which makes of use VLFeat MatConvNet library with different pretrained CNN models, including in particular the Caffe implementation of **AlexNet**[19] which is trained on the ImageNet dataset. The outputs from the 6th layer are used as *deep* features, leading to 4096 dimensional **DeCAF6** features. In this experiment, we denote the dataset **Amazon**,**Webcam**,**DSLR**,and **Caltech-256** as **A**,**W**,**D**,and **C**, respectively.

In denoting the direction from "source" to "target" by an arrow "→" is the direction from "source" to "target", $4 \times 3 = 12$ DA tasks can then be constructed, namely $A \rightarrow W \ldots C \rightarrow D$, respectively. For example, "W → D" means the Webcam image dataset is considered as the labeled *source* domain whereas the DSLR image dataset the unlabeled *target* domain.

USPS+MNIST shares ten common digit categories from two subsets, namely USPS and MNIST, but with very different data distributions (see Fig.2). We construct a first DA task *USPS vs MNIST* by randomly sampling 1,800 images in USPS to form the source data, and randomly sampling 2,000 images in MNIST to form the target data. Then, we switch the source/target pair to get another DA task, *i.e.*, *MNIST vs USPS*. We uniformly rescale all images to size 16×16, and represent each one by a feature vector encoding the gray-scale pixel values. Thus the source and target data share the same feature space. As a result, we have defined two cross-domain DA tasks, namely $USPS \rightarrow MNIST$ and $MNIST \rightarrow USPS$.

COIL20 contains 20 objects with 1440 images (Fig.2). The images of each object were taken in varying its pose about 5 degrees, resulting in 72 poses per object. Each image has a resolution of 32×32 pixels and 256 gray levels per pixel. In this experiment, we partition the dataset into two subsets, namely COIL 1 and COIL 2[42]. COIL 1 contains all images taken within the directions in $[0^0, 85^0] \cup [180^0, 265^0]$ (quadrants 1 and 3), resulting in 720 images. COIL 2 contains all images taken in the directions within $[90^0, 175^0] \cup [270^0, 355^0]$ (quadrants 2 and 4) and thus the number of images is 720. In this way, we construct two subsets with relatively different distributions. In this experiment, the COIL20 dataset with 20 classes is split into two DA tasks, *i.e.*, *COIL1 → COIL2* and *COIL2 → COIL1*

PIE face database consists of 68 subjects with each under 21 various illumination conditions[6], [22]. We adopt five pose subsets: C05, C07, C09, C27, C29, which provides a rich basis for domain adaptation, that is, we can choose one pose as the source and any rest one as the target. Therefore, we obtain $5 \times 4 = 20$ different source/target combinations. Finally, we combine all five poses together to form a single dataset for large-scale transfer learning experiment. We crop all images to 32×32 and only adopt the pixel values as the input. Finally, with different face poses, of which five subsets are selected, denoted as PIE1, PIE2, *etc*., resulting in $5 \times 4 = 20$ DA tasks, *i.e.*, *PIE1 vs PIE 2 ... PIE5 vs PIE 4*, respectively.

Dataset	DSLR	Amazon	Webcam	Caltech	MNIST	USPS	COIL1	COIL2	PIE1	PIE2	PIE3	PIE4	PIE5
No. Images	157	958	295	1123	2000	1800	720	720	3332	1629	1632	3329	1632
Classes	10	10	10	10	10	10	20	20	68	68	68	68	68
Feature Dimensions	Decaf(4096)/SURF(800)	Same	Same	Same	256	256	Pixel (1024)	same	Pixel (1024)	same	same	same	Same

Fig. 2: Sample images from six datasets used in our experiments. Each dataset represents a different domain. The OFFICE dataset contains three sub-datasets, namely DSLR, Amazon and Webcam.

B. Baseline Methods

The proposed DGA-DA method is compared with **twenty-two** methods of the literature, including deep learning-based approaches for unsupervised domain adaption. They are: (1)1-Nearest Neighbor Classifier(**NN**); (2) Principal Component

Analysis (**PCA**) +NN; (3) Geodesic Flow Kernel(**GFK**) [13] + NN; (4) Transfer Component Analysis(**TCA**) [28] +NN; (5) Transfer Subspace Learning(**TSL**) [36] +NN; (6) Joint Domain Adaptation (**JDA**) [22] +NN; (7) Extreme Learning Machine (**ELM**) [40] +NN; (8) Augmented Extreme Learning Machine (**AELM**) [40] +NN; (9) Subspace Alignment (**SA**)[9]; (10) Marginalized Stacked Denoising Auto-encoder (**mSDA**)[4]; (11) Transfer Joint Matching (**TJM**)[23]; (12) Robust Transfer Metric Learning (**RTML**)[6]; (13) Scatter Component Analysis (**SCA**)[11]; (14) Cross-Domain Metric Learning (**CDML**)[41]; (15)Deep Domain Confusion (**DDC**)[39]; (16)Low-Rank Transfer Subspace Learning (**LTSL**)[35]; (17)Low-Rank and Sparse Representation (**LRSR**)[42]; (18)Kernel Principal Component Analysis (**KPCA**)[34]; (19)Joint geometric and statistical alignment (**JGSA**) [44]; (20)Deep Adaptation Networks (**DAN**) [21]; (21)Deep Convolutional Neural Network (**AlexNet**) [19] and (22)Domain adaptation with low-rank reconstruction (**RVDLR**) [16].

In addition, for the purpose of fair comparison, we follow the experiment settings of **JGSA**, **AlexNet** and **SCA**, and apply DeCAF6 as the features for some methods to be evaluated. Whenever possible, the reported performance scores of the **twenty-two** methods of the literature are directly collected from previous research [22], [40], [6], [11], [42], [44]. They are assumed to be their *best* performance.

C. Experimental Setup

For the problem of domain adaptation, it is not possible to tune a set of optimal hyper-parameters, given the fact that the target domain has no labeled data. Following the setting of previous research[24], [22], [42] , we also evaluate the proposed **DGA-DA** by empirically searching in the parameter space for the *optimal* settings. Specifically, the proposed **DGA-DA** method has three hyper-parameters, *i.e.*, the subspace dimension k, regularization parameters λ and α. In our experiments, we set $k = 100$ and 1) $\lambda = 0.1$, and $\alpha = 0.99$ for **USPS**, **MNIST**，**COIL20** and **PIE**, 2) $\lambda = 1$, $\alpha = 0.99$ for **Office** and **Caltech-256**.

In our experiment, *accuracy* on the test dataset as defined by Eq.(24) is the evaluation measurement. It is widely used in literature, *e.g.*,[27], [21], [24], [22], [42], *etc.*

$$Accuracy = \frac{|x:x \in D_T \wedge \hat{y}(x) = y(x)|}{|x:x \in D_T|} \qquad (24)$$

where $\mathcal{D}_\mathcal{T}$ is the target domain treated as test data, $\hat{y}(x)$ is the predicted label and $y(x)$ is the ground truth label for a test data x.

To provide insight into the proposed DA method and highlight the individual contribution of each term in our final model, *i.e.*, the discriminative term using the repulsive force as defined in Eq.(12) and the geometry aware term through label smooth consistency as in Eq.(13) and geometry structure consistency as in Eq.(15), we evaluate the proposed DA method using three settings:

- **CDDA**: In this setting, sub-problem (b) in sect. III-C as defined in Eq.(19) is simply replaced by the Nearest Neighbor (NN) predictor. This correspond to our final DA model as defined in Eq.(18) which only makes use of the *repulse force* term but without geometry aware label inference as defined by Eq.(13) and Eq.(15). This setting makes it possible to understand how important discriminative DA is *w.r.t.* state of the art baseline DA methods only focused on data distribution alignment, *e.g.*, **JDA**.
- **GA-DA**: In this setting, we extend popular data distribution alignment-based DA methods, *e.g.*, **JDA**, with geometry aware label inference but ignore the *repulsive force* term, *i.e.*, $\max(\mathbf{A}^T \mathbf{X} \mathbf{M}_{\hat{c}} \mathbf{X}^T \mathbf{A})$, in our final model reformulated in Eq.(17). This setting thus jointly consider across domain conditional and marginal distribution alignment (Eq.(5) and Eq.(6)) and geometry aware label inference (Eq.(13) and Eq.(15)). This setting enables quantification of the contribution of the geometry aware label inference term as defined by Eq.(13) and Eq.(15) in comparison with state of the art baseline DA methods only focused on data distribution alignment, *e.g.*, **JDA**.
- **DGA-DA** : This setting correspond to our full final model as defined in Eq.(17). It thus contains **CDDA** as expressed by sub-problem (a) as in sect. III-C to which we further add the geometry aware label inference as defined by sub-problem (b) in sect. III-C.

D. Experimental Results and Discussion

*1) **Experiments on the COIL 20 Dataset**:* The COIL dataset (see fig.2) features the challenge of pose variations between the source and target domain. Fig.3 depicts the experimental results on the COIL dataset. As can be seen in this figure where top results are highlighted in red color, the two partial models, *i.e.*, **CDDA**, **DA-GA** and the proposed final model, **DGA-DA**, depict an overall average accuracy of 92.71%, 90.70% and 100.00%, respectively. They both outperform the eight baseline DA algorithms with a significant margin.

It is worth noting that, when adding label inference based on the underlying data manifold structure, the proposed **DGA-DA** improves its sibling **CDDA** by a margin as high as roughly 7 points, thereby highlighting the importance of data geometry aware label inference as introduced in **DGA-DA**. As compared to **JDA**, the proposed **CDDA**, which adds a discriminative *repulsive force* term *w.r.t.* **JDA**, also shows its effectiveness and improves the latter by more than 3 points.

	COIL1 → COIL2	COIL2 → COIL1	Average
■ NN	83.61	82.78	83.20
■ PCA	84.72	84.03	84.38
■ GFK	72.50	74.17	73.34
■ TSL	88.06	87.92	87.99
■ LTSL	75.69	72.22	73.96
■ LRSR	88.61	89.17	88.89
■ TCA	88.47	85.83	87.15
■ JDA	89.31	88.47	88.89
■ CDDA	91.53	93.89	92.71
■ GA-DA	89.86	91.53	90.70
■ DGA-DA	100	100	100

Fig. 3: Accuracy% on the COIL Images Dataset.

	C→A	C→W	C→D	A→C	A→W	A→D	W→C	W→A	W→D	D→C	D→A	D→W	Average
── PCA	85.60	66.10	74.50	70.30	57.20	64.90	60.30	62.50	98.70	52.00	62.70	89.10	70.40
── NN	87.05	72.20	80.89	78.54	77.31	80.25	68.21	73.07	100.00	70.08	75.89	97.97	80.12
── ELM	89.07	70.51	78.98	79.61	74.58	80.25	70.61	75.37	100.00	68.21	80.79	98.31	80.52
── GFK	87.27	75.93	83.44	80.32	76.95	80.89	67.76	74.32	100.00	69.10	75.78	98.64	80.87
── SA	87.06	75.59	80.25	79.61	78.31	81.53	68.83	75.16	100.00	69.99	73.49	98.98	80.73
── mSDA	89.67	68.47	82.17	78.81	78.98	79.62	69.46	76.62	100.00	73.29	81.32	98.64	81.42
── TJM	88.10	72.20	74.52	77.65	75.25	82.80	71.42	80.27	100.00	72.57	78.60	98.31	80.97
── AELM	89.46	79.32	81.53	79.96	77.63	85.35	71.24	76.83	100.00	75.60	83.19	98.98	83.25
── RTML	90.20	83.80	88.70	83.10	79.50	83.80	82.90	90.80	100.00	81.60	90.60	98.60	87.80
── SCA	89.46	85.42	87.90	78.81	75.93	85.35	74.80	86.12	100.00	78.09	89.98	98.64	85.88
── JGSA	91.44	86.78	93.63	84.86	81.02	88.54	84.95	90.71	100.00	86.20	91.96	99.66	89.98
── AlexNet	91.90	83.70	87.10	83.00	79.50	87.40	73.00	83.80	100.00	79.00	87.10	97.70	86.10
── DAN	92.00	90.60	89.30	84.10	91.80	91.70	81.20	92.10	100.00	80.30	90.00	98.50	90.10
── DDC	91.90	85.40	88.80	85.00	86.10	89.00	78.00	84.90	100.00	81.10	89.50	98.20	88.20
── JDA	89.70	83.70	86.60	82.20	78.60	80.20	80.50	88.10	100.00	80.10	89.40	98.90	86.50
─ ─ ─ CDDA	90.71	85.76	91.72	85.66	78.31	84.08	86.02	89.77	100.00	86.20	91.34	100.00	89.13
─ ─ ─ GA-DA	90.65	87.80	94.27	84.51	82.03	86.62	84.95	91.44	100.00	85.75	93.53	99.66	90.10
── DGA-DA	91.25	93.56	91.72	85.20	80.98	89.81	86.46	90.81	100.00	86.20	93.11	100.00	90.76

Fig. 4: Accuracy% on the Office+Caltech Images with DeCAF6 Features.

*2) **Experiments on the Office+Caltech-256 Data Sets**:* Fig.4 and Fig.5 synthesize the experimental results in comparison with the state of the art when deep features (*i.e.*, DeCAF6 features) and classic shallow features (*i.e.*, SURE features) are used, respectively.

- As can be seen in Fig.5, both **CDDA** and **DGA-DA** outperform the state of the art method in terms of average accuracy, thereby demonstrating the effectiveness of the proposed DA method. In particular, in comparison with **JDA** which only cares about data distribution alignment between source and target and the proposed DA method is built upon, **CDDA** improves **JDA** by 2 points thanks to the discriminative repulsive force term introduced in our model. When label inference accounts for the underlying data structure, our final model **DGA-DA** further improves **CDDA** by roughly 1 point.

- Fig.4 compares the proposed DA method using deep features *w.r.t.* the state of the art, in particular end-to-end deep learning-based DA methods. As can be seen in Fig.4, the use of deep features has enabled impressive accuracy improvement over shallow features. Simple baseline methods, *e.g.*, **NN**, **PCA**, see their accuracy soared by roughly 40 points, demonstrating the power of deep learning paradigm. Our proposed DA method also takes advantage of this jump and sees its accuracy soared from 48.22 to 89.13 for **CDDA** and from 49.02 to 90.43 for **DGA-DA**. As for shallow features, **CDDA** improves **JDA** by 3 points and **DGA-DA** further ameliorates **CDDA** by 1 point when label inference accounts for the underlying data geometric structure. As a result, **DGA-DA** displays the best average accuracy and outperforms slightly **DAN**.

*3) **Experiments on the USPS+MNIST Data Set**:* The UPS+MNIST dataset features different writing styles between source and target. Fig.6 lists the experimental results in comparison with 14 state of the art DA methods. As can be seen in the table, **CDDA** displays a 69.14% average accuracy and ranks the third best performer. It shows its effectiveness once more as it improves its baseline **JDA** by more than 5 points on average. When accounting for the underlying data geometry structure, the proposed **DGA-DA** further improves its sibling **CDDA** by a margin more than 7 points and displays the state of the art performance of a 76.54% accuracy. It is worth noting that the second best DA performer on this dataset, *i.e.*, **JGSA**, also suggests aligning both statistically and geometrically data, and thereby corroborates our data geometry aware DA approach.

*4) **Experiments on the CMU PIE Data Set**:* The CMU PIE dataset is a large face dataset featuring both illumination and pose variations. Fig.7 synthesizes the experimental results for DA using this dataset. As can be seen in the figure, similarly as in the previous experiments, the proposed **DGA-DA** displays the best average accuracy over 20 cross-domain adaptation experiments. In aligning both marginal and conditional data distributions, **JDA** performs quite well and displays a 60.24%

	C→A	C→W	C→D	A→C	A→W	A→D	W→C	W→A	W→D	D→C	D→A	D→W	Average
■NN	23.70	25.76	25.48	26.00	29.83	25.48	19.86	22.96	59.24	26.27	28.50	63.39	31.37
■PCA	36.95	32.54	38.22	34.73	35.59	27.39	26.36	31.00	77.07	29.65	32.05	75.93	39.79
■GFK	41.02	40.68	38.85	40.25	38.98	36.31	30.72	29.75	80.89	30.28	32.05	75.59	42.95
■KPCA	40.40	31.53	40.76	37.04	31.86	33.76	27.60	29.44	89.81	27.78	31.00	84.41	42.12
■SCA	43.74	33.56	39.49	38.29	33.90	34.21	30.63	30.48	92.36	32.32	33.72	88.81	44.29
■LTSL	25.26	19.32	21.02	16.92	14.58	21.02	34.64	39.56	72.61	35.08	39.67	74.92	34.55
■LRSR	51.25	38.64	47.13	43.37	36.61	38.85	29.83	34.13	82.80	31.61	33.19	77.29	45.39
■CDML	47.70	35.60	42.50	40.70	37.30	35.30	31.60	32.40	77.90	32.20	29.40	79.40	43.50
■mSDA	45.92	37.96	46.49	40.96	40.33	36.30	31.96	33.61	87.26	30.89	35.59	87.45	46.23
■ELM	49.37	37.79	45.22	40.07	33.56	34.31	31.17	33.85	88.54	28.23	28.50	73.22	43.65
■AELM	53.13	49.49	50.96	41.14	35.25	36.94	34.11	38.93	89.81	33.83	33.09	80.33	48.08
■SA	41.02	40.34	47.13	40.16	39.66	35.03	31.17	33.82	85.99	31.26	35.80	84.75	45.51
■TJM	46.76	38.98	44.59	39.45	42.03	45.22	30.19	29.96	89.17	31.43	32.78	85.42	46.33
■TSL	44.47	34.24	43.31	37.58	33.90	26.11	29.83	30.27	87.26	28.50	27.56	85.42	42.37
■TCA	38.20	38.64	41.40	37.76	37.63	33.12	29.30	30.06	87.26	31.70	32.15	86.10	43.61
■JDA	44.78	41.69	45.22	39.36	37.97	39.49	31.17	32.78	89.17	31.52	33.09	89.49	46.31
■CDDA	48.33	44.75	48.41	42.12	41.69	37.58	31.97	37.27	87.90	34.64	33.51	90.51	48.22
■GA-DA	48.96	44.41	47.13	39.09	44.07	37.58	22.89	29.13	89.81	26.45	36.53	92.54	46.55
■DGA-DA	52.09	47.12	45.86	41.32	38.31	38.22	33.30	41.75	89.81	33.66	33.61	93.22	49.02

Fig. 5: Accuracy% on the Office+Caltech Images with SURF-BoW Features.

	USPS→MNIST	MNIST→USPS	Average
■NN	44.70	65.94	55.32
■PCA	44.95	66.22	55.59
■GFK	46.45	67.22	56.84
■TSL	53.75	66.06	59.91
■KPCA	42.55	62.61	52.58
■SCA	48.00	65.11	56.56
■TJM	52.52	63.28	57.90
■ELM	57.70	61.11	59.41
■SA	40.15	48.22	44.19
■mSDA	43.20	66.94	55.07
■AELM	57.77	62.33	60.05
■JGSA	68.15	80.44	74.30
■TCA	51.05	56.28	53.67
■JDA	59.65	67.28	63.47
■CDDA	62.05	76.22	69.14
■GA-DA	59.80	75.39	67.60
■DGA-DA	70.75	82.33	76.54

Fig. 6: Accuracy% on the USPS+MNIST Images Dataset.

average accuracy. In integrating the discriminative repulsive force term, **CDDA** improves **JDA** by roughly 3 points. **DGA-DA** further ameliorates **CDDA** by more than 1 point.

It is interesting to note that the second best performer on this dataset, namely **LRSR**, also tries to align geometrically source and target data through both low rank and sparse constraints so that source and target data are interleaved within a novel shared feature subspace.

	PIE 1 5→7	PIE 2 5→9	PIE 3 5→27	PIE 4 5→29	PIE 5 7→5	PIE 6 7→9	PIE 7 7→27	PIE 8 7→29	PIE 9 9→5	PIE 10 9→7	PIE 11 9→27	PIE 12 9→29	PIE 13 27→5	PIE 14 27→7	PIE 15 27→9	PIE 16 27→29	PIE 17 29→5	PIE 18 29→7	PIE 19 29→9	PIE 20 29→27	Average
■NN	26.09	26.59	30.67	16.67	24.49	46.63	54.07	26.53	21.37	41.01	46.53	26.23	32.95	62.68	73.22	37.19	18.49	24.19	28.31	31.24	34.76
■PCA	24.80	25.18	29.26	16.30	24.22	45.53	53.35	25.43	20.95	40.45	46.14	25.31	31.96	60.96	72.18	35.11	18.85	23.39	27.21	30.34	33.85
■GFK	26.15	27.27	31.15	17.59	25.24	47.37	54.25	27.08	21.82	43.16	46.41	26.78	34.24	62.92	73.35	37.38	20.35	24.62	28.49	31.33	35.35
■CDML	53.22	53.12	80.12	48.23	52.39	54.23	68.36	37.34	43.54	54.87	62.76	38.21	75.12	80.53	83.72	52.78	27.34	30.82	36.34	40.61	53.69
■RTML	60.12	55.21	85.19	52.98	58.13	63.92	76.16	40.38	53.12	58.67	69.81	42.13	81.12	83.92	89.51	56.26	29.11	33.28	39.85	47.13	58.80
■LTSL	22.96	20.65	31.81	12.07	18.25	16.05	45.15	17.52	22.36	20.26	57.34	24.57	51.20	70.10	72.00	48.28	13.06	21.61	17.03	29.59	31.59
■mSDA	28.35	26.91	30.39	21.76	28.27	44.19	55.39	28.08	24.83	42.59	50.25	27.83	32.89	63.01	74.70	34.81	25.85	26.33	28.63	32.98	36.41
■RDALR	40.76	41.79	59.63	29.35	41.81	51.47	64.73	33.70	34.69	47.70	56.23	33.15	55.64	67.83	75.86	40.26	26.98	29.90	29.90	33.64	44.75
■LRSR	65.87	64.09	82.03	54.90	45.04	53.49	71.43	47.97	52.49	55.56	77.50	54.11	81.54	85.39	82.23	72.61	52.19	49.41	58.45	64.31	63.53
■TSL	44.08	47.49	62.78	36.15	46.28	57.60	71.43	35.66	36.94	47.02	59.45	36.34	63.66	72.68	83.52	44.79	33.28	34.13	36.58	38.75	49.43
■TCA	40.76	41.79	59.63	29.35	41.81	51.47	64.73	33.70	34.69	47.70	56.23	33.15	55.64	67.83	75.86	40.26	26.98	29.90	29.90	33.64	44.75
■JDA	58.81	54.23	84.50	49.75	57.62	62.93	75.82	39.89	50.96	57.95	68.45	39.95	80.58	82.63	87.25	54.66	46.46	42.05	53.31	57.01	60.24
■CDDA	60.22	58.70	83.48	54.17	62.33	64.64	79.90	44.00	58.46	59.73	77.20	47.24	83.10	82.26	86.64	58.33	48.02	45.61	52.02	55.99	63.10
■GA-DA	57.40	60.54	84.05	52.21	57.89	61.58	82.34	41.42	54.14	60.77	77.23	43.50	79.83	84.71	89.17	53.62	52.73	47.64	51.66	58.82	62.56
■DGA-DA	65.32	62.81	83.54	56.07	63.69	61.27	82.37	46.63	56.72	61.26	77.83	44.24	81.84	85.27	90.95	53.80	57.44	53.84	55.27	61.82	65.10

Fig. 7: Accuracy% on the PIE Images Dataset.

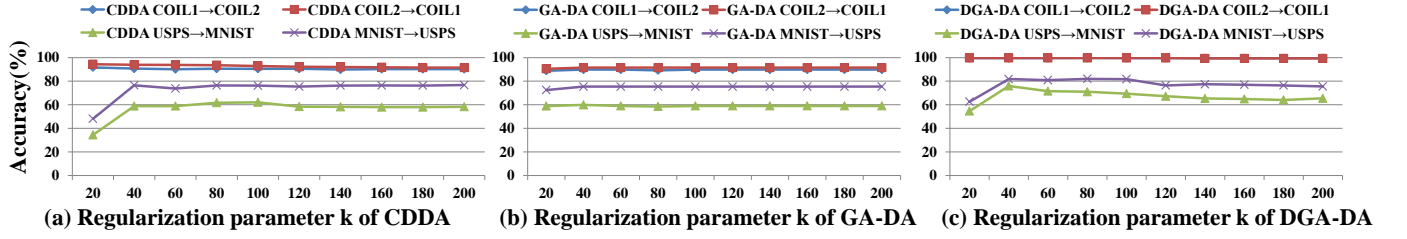## E. Convergence and Parameter Sensitivity

Fig. 8: Sensitivity analysis of the proposed methods: (a) accuracy *w.r.t.* subspace dimension k of CDDA; (b)accuracy *w.r.t.* subspace dimension k of GA-DA; (c) accuracy *w.r.t.* subspace dimension k of DGA-DA. Four datasets are used, *i.e.*, COIL1, COIL2, USPS and MNIST.
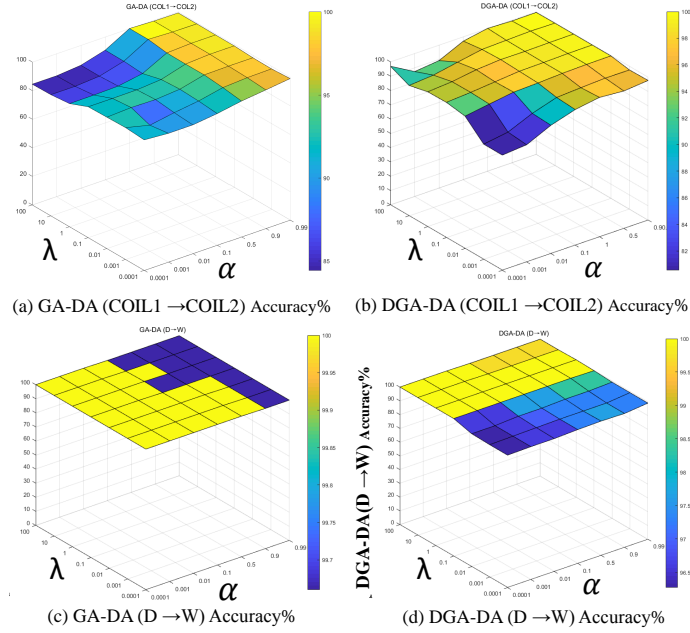

(a) GA-DA (COIL1 →COIL2) Accuracy%   (b) DGA-DA (COIL1 →COIL2) Accuracy%

(c) GA-DA (D →W) Accuracy%   (d) DGA-DA (D →W) Accuracy%

Fig. 9: The classification accuracies of the proposed **GA-DA** and **DGA-DA** method vs. the parameters α and λ on the selected four cross domains data sets, *i.e.*, DSLR (D), Webcam (W), COIL1 and COIL2, with k held fixed at 100.

Fig. 10: Convergence analysis using 12 cross-domain image classification tasks on Office+Caltech256 datasets with DeCAF6 Features. (accuracy w.r.t #iterations)

While the proposed **DGA-DA** displays state of the art performance over 36 DA tasks through six datasets (USPS, MINIST, COIL20, PIE, Amazon, Caltech), an important question is how fast the proposed method converges (sect.IV-E2) as well as its sensitivity *w.r.t.* its hyper-parameters (Sect.IV-E1).

Fig. 11: Comparisons of baseline domain adaptation methods and the proposed **CDDA**, **GA-DA** and **DGA-DA** method on the synthetic data
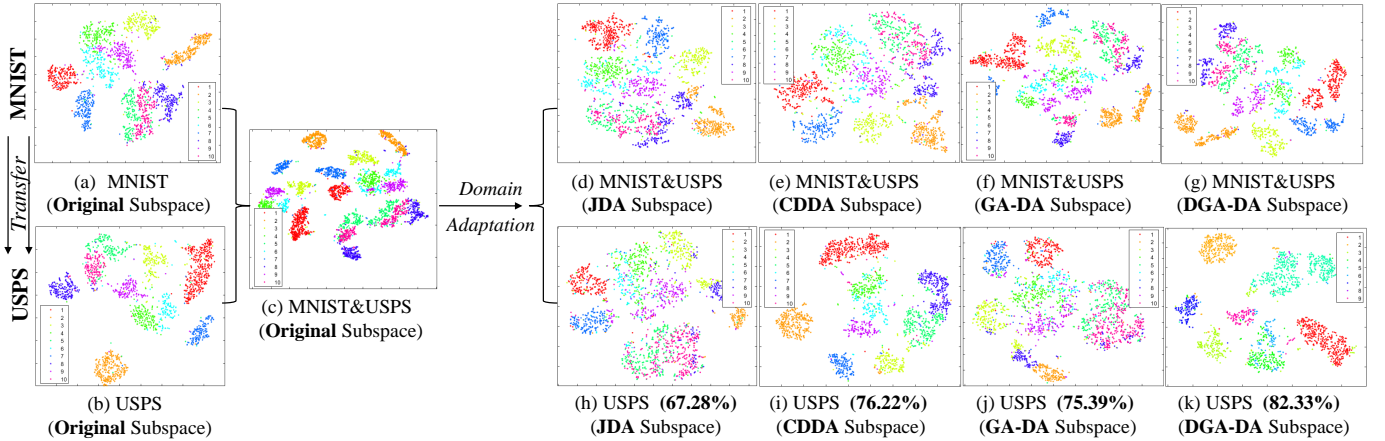

Fig. 12: Accuracy(%) and Visualization results of the MNIST→USPS DA task. Fig.12(a), Fig.12(b) and Fig.12(c) are visualization results of MNIST, USPS, MNIST&USPS datasets in their **Original** data space, respectively. After domain adaptation, Fig.12(d), Fig.12(e), Fig.12(f) and Fig.12(g) visualize the MNIST&USPS datasets in **JDA**, **CDDA**, **GA-DA** and **DGA-DA** subspaces, respectively. Fig.12(h), Fig.12(i), Fig.12(j) and Fig.12(k) show the visualization results of the target domain USPS in **JDA**, **CDDA**, **GA-DA** and **DGA-DA** subspaces, respectively. The ten digit classes are represented by different colors.

1) Parameter sensitivity: Three hyper-parameters, namely k, λ and α, are introduced in the proposed methods. k is the dimension of the extracted feature subspace which determines the structure of low-dimension embedding. In Fig.8, we plot the classification accuracies of the proposed DA method *w.r.t* different values of k on the **COIL** and **USPS+MINIST** datasets. As shown in Fig.8, the subspace dimensionality k varies with $k \in \{20, 40, 60, 80, 100, 120, 140, 160, 180, 200\}$, yet the proposed 3 DA variants, namely, **CDDA**, **GA-DA** and **DGA-DA**, remain stable *w.r.t.* a wide range of with $k \in \{40 \leq k \leq 200\}$. In our experiments, we set $k = 100$ to balance efficiency and accuracy.

λ as introduced in Eq.(17) and Eq.(18) aims to regularize the projection matrix **A** to avoid over-fitting the chosen shared feature subspace with respect to both source and target data. $\alpha = \frac{1}{1+\mu}$ as defined in Eq.(22) is a trade-off parameter which balances LSC and GSC. We study the sensitivity of the proposed **GA-DA** and **DGA-DA** methods with a wide range of parameter values, *i.e.*, $\alpha = (0.0001, 0.001, 0.01, 0.1, 0.5, 0.9, 0.99)$ and $\lambda = (0.0001, 0.001, 0.01, 0.1, 1, 10, 100)$. We plot in

Fig.9 the results on $D \rightarrow W$ and $COIL1 \rightarrow COIL2$ datasets on both methods with k held fixed at 100. As can be seen from Fig.9, the proposed **GA-DA** and **DGA-DA** display their stability as the resultant classification accuracies remain roughly the same despite a wide range of λ and α values.

2) Convergence analysis: In Fig.10, we further perform convergence analysis of the proposed **CDDA**, **GA-DA** and **DGA-DA** methods using the **DeCAF6** features on the **Office+Caltech** datasets. The question here is how fast a DA method achieves its best performance *w.r.t.* the number of iterations T. Fig.10 reports 12 cross domain adaptation experiments ($C \rightarrow A$, $C \rightarrow W$ ... $D \rightarrow A$, $D \rightarrow W$) with the number of iterations $T = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$.

As shown in Fig.10, **CDDA**, **GA-DA** and **DGA-DA** converge within 3∼5 iterations during optimization.

F. Analysis and Verification

To further gain insight of the proposed **CDDA**, **GA-DA** and **DGA-DA** *w.r.t.* its domain adaptation skills, we also evaluate the proposed methods using a synthetic dataset in comparison with several state of the art DA methods. Fig.11 visualizes the original data distributions with 4 classes and the resultant shared feature subspaces as computed by **TCA**, **JDA**, **TJM**, **SCA**, **CDDA**, **GA-DA** and **DGA-DA**, respectively. In this experiment, we focus our attention on the ability of the DA methods to: : (a) narrow the discrepancies of data distributions between source and target; (b) increase data discriminativeness; and (c) align data geometric structures between source and target. As such, the original synthetic data depicts slight distribution discrepancies between source and target for the first two class data, wide distribution mismatch for the third and fourth class data. Fourth class data further depict a moon like geometric structure.

As can be seen in Fig.11, baseline methods, *e.g.*, **TCA**, **SCA**, **TJM** have difficulties to align data distributions with wide discrepancies, *e.g.*, third class data. **JDA** narrow data distribution discrepancies but lacks class data discriminativeness. The proposed variant **CDDA** ameliorates **JDA** and makes class data well separated thanks to the introduced *repulsive force* term but falls short to preserve data geometric structure (see the fourth moon like class data. The variant **GA-DA** align data distributions and preserves the underlying data geometric structures thanks to label smoothness consistency (LSC) and geometric structure consistency (GSC) but lacks data discriminativeness. In contrast, thanks to the joint consideration of data discriminativeness and geometric structure awareness, the proposed **DGA-DA** not only align data distributions compactly but also separate class data very distinctively. Furthermore, it also preserves the underlying data geometric structures.

The above findings can be further verified using real data through the MNIST→USPS DA task where the proposed DA methods achieves remarkable results (See Fig.6). Fig.12 visualizes class explicit data distributions in their original subspace and the resultant shared feature subspace using **JDA** and the three variants of the proposed DA method, namely **CDDA**, **GA-DA** , **DGA-DA**, with the same experimental setting.

- Data distributions and geometric structures. Fig.12(a,b,c) visualize the MNIST, USPS, MNIST&USPS datasets in their **Original** data space, respectively. As shown in these figures, the MNIST and USPS datasets depict different data distributions and various data structures. In particular, yellow dots represent digit 2. They show a long and narrow shape in MNIST (Fig.12(a)) while a circle like shape in USPS (Fig.12(b)). They further display large data discrepancies across domain (Fig.12(c)) as for all the other classes.

- Contribution of the *repulsive force* term. Visualization results in Fig.12(h,i,j,k) show that, in comparison with their respective baseline DA methods, *i.e.*, **JDA** (Fig.12(h)) and **GA-DA** (Fig.12(j)), the proposed two DA variants, *i.e.*, **CDDA**(Fig.12(i)) and **DGA-DA**(Fig.12(k)) which integrate in their model the *repulsive force* term as introduced in Sect.III-B3, achieve data discriminativeness in compacting intra-class instances and separating inter-class data, respectively. As a result, as shown in Fig.6, **DGA-DA** outperforms **GA-DA** by $6.94 \uparrow$ points, and **CDDA** outperforms **JDA** by $8.94 \uparrow$ points,respectively, thereby illustrating the importance of increasing data discriminativeness in DA.

- Contribution of Geometric Structure Awareness. Visualization results in Fig.12(d,e) show that the **JDA** and **CDDA**'s subspaces fail to preserve the geometric structures of the underlying data manifold. For instance, the long and narrow shape of the orange dots in the source MNIST domain and the corresponding circle blob orange cloud in the target USPS domain (Fig.12(c)) are not preserved anymore in the **JDA** (Fig.12(d)) and **CDDA** (Fig.12(e)) subspaces. In contrast, thanks to the geometry awareness constraints, *i.e.*, label smoothness consistency (LSC) and geometric structure consistency (GSC), as introduced in Sect.III-B4, the two variants of the proposed DA methods, *i.e.*, **DA-GA** (Fig.12(f)) and **DGA-DA** (Fig.12(g)), succeed to preserve the geometric structures of the underlying data, and thereby inherent data similarities and consistencies of label inference. As a result, **DGA-DA** outperforms **CDDA** by $6.11 \uparrow$ points, and **GA-DA** outperforms **JDA** by $8.11 \uparrow$ points. They thus suggest the importance of Geometric Structure Awareness in DA.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel Discriminative and Geometry Aware Unsupervised DA method based on feature adaptation. Comprehensive experiments on 36 cross-domain image classification tasks through six popular DA datasets highlight the interest of enhancing the data discriminative properties within the model and label propagation in respect of the geometric structure of the underlying data manifold, and verify the effectiveness of the proposed method compared with twenty-two baseline DA methods of the literature. Using both synthetic and real data and three variants of the proposed DA method, we

have further provided in-depth analysis and insights into the proposed **DGA-DA**, in quantifying and visualizing the contribution of the data discriminativeness and data geometry awareness.

Our future work will concentrate on embedding the proposed method in deep networks and study other vision tasks, *e.g.*, object detection, within the setting of transfer learning. Our future work will concentrate on embedding the proposed method in deep networks and study other vision tasks, *e.g.*, object detection, within the setting of transfer learning.

REFERENCES

[1] Mahsa Baktashmotlagh, Mehrtash Harandi, and Mathieu Salzmann. Distribution-matching embedding for visual domain adaptation. *Journal of Machine Learning Research*, 17(108):1–30, 2016. 3

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 1

[3] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 3, 4, 5

[4] Minmin Chen, Zhixiang Eddie Xu, Kilian Q. Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. *CoRR*, abs/1206.4683, 2012. 10

[5] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *CoRR*, abs/1702.05374, 2017. 1

[6] Zhengming Ding and Yun Fu. Robust transfer metric learning for image classification. *IEEE Trans. Image Processing*, 26(2):660–670, 2017. 9, 10

[7] Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Semi-supervised domain adaptation with instance constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 668–675, 2013. 1

[8] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 647–655, 2014. 9

[9] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 2960–2967, 2013. 10

[10] Michel Fortin and Roland Glowinski. *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*, volume 15. Elsevier, 2000. 7, 8

[11] Muhammad Ghifary, David Balduzzi, W. Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(7):1414–1430, 2017. 3, 9, 10

[12] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 222–230, 2013. 9

[13] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012. 1, 9, 10

[14] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012. 1

[15] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(5):550–554, 1994. 9

[16] I-Hong Jhuo, Dong Liu, DT Lee, and Shih-Fu Chang. Robust visual domain adaptation with low-rank reconstruction. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2168–2175. IEEE, 2012. 10

[17] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 180–191. VLDB Endowment, 2004. 1, 2

[18] T. H. Kim, K. M. Lee, and S. U. Lee. Learning full pairwise affinities for spectral segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1690–1703, July 2013. 6, 7

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 9, 10

[20] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. 9

[21] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015. 5, 10

[22] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2200–2207, 2013. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

[23] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer joint matching for unsupervised domain adaptation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1410–1417, 2014. 3, 10

[24] Lingkun Luo, Xiaofang Wang, Shiqiang Hu, and Liming Chen. Robust data geometric structure aligned close yet discriminative domain adaptation. *CoRR*, abs/1705.08620, 2017. 3, 4, 9, 10

[25] Lingkun Luo, Xiaofang Wang, Shiqiang Hu, Chao Wang, Yuxing Tang, and Liming Chen. Close yet distinctive domain adaptation. *CoRR*, abs/1704.04235, 2017. 3

[26] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2002. 7, 8

[27] Sinno Jialin Pan, James T Kwok, and Qiang Yang. Transfer learning via dimensionality reduction. In *AAAI*, volume 8, pages 677–682, 2008. 3, 6, 10

[28] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. 2, 3, 4, 6, 10

[29] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 1

[30] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[31] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, May 2015. 1

[32] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2988–2997, 2017. 1

[33] Sandeepkumar Satpal and Sunita Sarawagi. Domain adaptation of conditional probability models via feature subsetting. In *PKDD*, volume 4702, pages 224–235. Springer, 2007. 3

[34] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. 10

[35] Ming Shao, Dmitry Kit, and Yun Fu. Generalized transfer subspace learning through low-rank constraint. *International Journal of Computer Vision*, 109(1-2):74–93, 2014. 1, 3, 4, 9, 10

[36] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, July 2010. 1, 3, 10

[37] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008. 3

[38] Yuxing Tang, Josiah Wang, Xiaofang Wang, Boyang Gao, Emmanuel Dellandrea, Robert Gaizauskas, and Liming Chen. Visual and semantic knowledge transfer for large scale semi-supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1

[39] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014. 10

[40] Muhammad Uzair and Ajmal S. Mian. Blind domain adaptation with augmented extreme learning machine features. *IEEE Trans. Cybernetics*, 47(3):651–660, 2017. 9, 10

[41] Hao Wang, Wei Wang, Chen Zhang, and Fanjiang Xu. Cross-domain metric learning based on information theory. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 2099–2105, 2014. 10

[42] Yong Xu, Xiaozhao Fang, Jian Wu, Xuelong Li, and David Zhang. Discriminative transfer subspace learning via low-rank and sparse representation. *IEEE Trans. Image Processing*, 25(2):850–863, 2016. 1, 3, 6, 9, 10

[43] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang. Saliency detection via graph-based manifold ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, June 2013. 6, 7

[44] Jing Zhang, Wanqing Li, and Philip Ogunbona. Joint geometrical and statistical alignment for visual domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3, 4, 10

[45] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schlkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pages 321–328. MIT Press, 2004. 3, 7

Lingkun Luo received his first master's degree in computer science from Hosei University and second master's degree in software engineering from University of Science and Technology of China. Now, he is a PhD candidate at Shanghai Jiao Tong University (SJTU). He is currently a research assistant in Ecole Centrale de Lyon (ECL), Department of Mathematics and Computer Science, and a member of LIRIS laboratory. He is jointly supervised by SJTU and ECL. His research interests include machine learning, pattern recognition and computer vision.
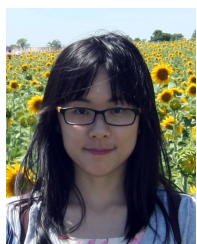
Liming Chen received the joint B.Sc. degree in mathematics and computer science from the University of Nantes, Nantes, France in 1984, and the M.Sc. and Ph.D. degrees in computer science from the University of Paris 6, Paris, France, in 1986 and 1989, respectively.

He first served as an Associate Professor with the Université de Technologie de Compiègne, before joining École Centrale de Lyon, Écully, France, as a Professor in 1998, where he leads an advanced research team on multimedia computing and pattern recognition. From 2001 to 2003, he also served as Chief Scientific Officer in a Paris-based company, Avivias, specializing in media asset management. In 2005, he served as Scientific Multimedia Expert for France Telecom R&D China, Beijing, China. He was the Head of the Department of Mathematics and Computer Science, École Centrale de Lyon from 2007 through 2016. His current research interests include computer vision, machine learning, image and video analysis and categorization, face analysis and recognition, and affective computing. Liming has over 250 publications and successfully supervised over 35 PhD students. He has been a grant holder for a number of research grants from EU FP program, French research funding bodies and local government departments. Liming has so far guest-edited 3 journal special issues. He is an associate editor for Eurasip Journal on Image and Video Processing and a senior IEEE member.

Shiqiang Hu received his PhD degree at Beijing Institute of Technology. His research interests include data fusion technology, image understanding, and nonlinear filter.

Ying LU received the B.S. degree in Applied Mathematics and the M.S. degree in Computer Science and Engineering from Beihang University, Beijing, China, in 2010 and 2013, respectively, and Ph.D. degree in computer science from University of Lyon, France, in 2017. She also received a Research Master's degree in Computer Science from Ecole Centrale de Lyon, University Lyon I and INSA Lyon in 2012, and an Engineering degree from Ecole Centrale de P é kin, Beihang Universy in 2013. She is currently a teaching and research assistant in Ecole Centrale de Lyon, Department of Mathematics and Computer Science, and a member of LIRIS laboratory. Her research interests include machine learning and computer vision.

Xiaofang Wang is currently assistant lecturer and researcher in Ecole Centrale Lyon. She has received the B.S. and M.S. degrees in biomedical engineering from Central South University, Changsha, China, and the Ph.D. degree in computer science from École Centrale de Lyon, France in 2015.

Her current research interests include image/video processing, machine learning (transfer learning, deep learning), computer vision (semantic image segmentation, object localization and recognition, etc.).