Discriminative Noise Robust Sparse Orthogonal Label Regression-based Domain Adaptation

Lingkun Luo, Liming Chen, Senior Member, IEEE, and Shiqiang Hu,

Abstract—Domain adaptation (DA) aims to enable a learning model trained from a source domain to generalize well on a target domain, despite the mismatch of data distributions between the two domains. State-of-the-art DA methods have so far focused on the search of a latent shared feature space where source and target domain data can be aligned either statistically and/or geometrically. In this paper, we propose a novel unsupervised DA method, namely Discriminative Noise Robust Sparse Orthogonal Label Regression-based Domain Adaptation (DOLL-DA). The proposed DOLL-DA derives from a novel integrated model which searches a shared feature subspace where source and target data are, through optimization of some repulse force terms, discriminatively aligned statistically, while at same time regresses orthogonally data labels thereof using a label embedding trick. Furthermore, in minimizing a novel Noise Robust Sparse Orthogonal Label Regression (NRS-OLR) term, the proposed model explicitly accounts for data outliers to avoid negative transfer and introduces the property of sparsity when regressing data labels. We carry out comprehensive experiments in comparison with 32 state of the art DA methods using 8 standard DA benchmarks and 49 cross-domain image classification tasks. The instance-based approach generally assumes that 1) the conditional distributions of source and target domain are identical[64], and 2) certain portion of the data in the source domain can be reused[43] for learning in the target domain through re-weighting. Feature-based adaptation[54], [21], [14], [37], [39] 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 novel hybrid DA method using both feature adaptation and classifier optimization.

I. INTRODUCTION

Traditional machine learning tasks assume that both training and testing data are drawn from the same data distribution [43], [45]. 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 with respect to the labeled data in a source domain where a predictor can not be reliably learned. On the other hand, manually labeling enough target domain data for the purpose of training an effective predictor can be very expensive, tedious and thus prohibitive.

Domain adaptation (DA) [43], [45], [37] aims to leverage possibly abundant labeled data from a source domain to learn an effective predictor for data in an unseen domain, namely target domain, despite the data distribution discrepancy between the source and target. While DA can be semi-supervised by assuming that a certain amount of labeled data is available in the target domain, in this paper we are interested in unsupervised DA where we assume that no labels are available for target domain data.

While there exists an increasing number of deep learning-based unsupervised DA methods, we focus in this paper on shallow DA methods as they are easier to train and can provide insights into the design decisions of deep DA methods. The relationships between shallow and deep DA methods will be discussed in depth in Sect.II on related works. State of the art shallow DA methods can be categorized into instance-based [43], [10], feature-based [44], [33], [63], or classifier-based. Classifier-based DA is widely applied in semi-supervised DA as it aims to fit a classifier trained on source domain data to target domain data through adaptation of its parameters, and thereby require some labels in the target domain[56].

The instance-based approach generally assumes that 1) the conditional distributions of source and target domain are identical[64], and 2) certain portion of the data in the source domain can be reused[43] for learning in the target domain through re-weighting. Feature-based adaptation[54], [21], [14], [37], [39] 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 novel hybrid DA method using both feature adaptation and classifier optimization.

A common method to approach feature adaptation is to seek a shared latent subspace between the source and target domain [45], [44] via optimization. State-of-the-art features three main lines of approaches, namely, data geometric structure alignment-based (DGSA), data distribution centered (DDC) or their hybridization. DGSA based approaches [54], [51], [63] 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. DDC methods [38], [35], [30], [36] 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[53], Geodesic distance[16], Wasserstein
Our previously proposed DA method, namely DGA-DA[39], is a hybridization of DDC and DGSA approaches. DGA-DA leverages the advantages of both DDC and DGSA methods and demonstrates a state of the art performance on a number of DA benchmarks. DGA-DA relies upon the analysis of a cornerstone theoretical result in DA [2], [1], [26], which estimates an error bound of a learned hypothesis \( h \) on a target domain as follows:

\[
e_{T}(h) \leq e_{S}(h) + d_{H}(D_{S}, D_{T}) + \min \{ \mathcal{E}_{D_{S}}[\|f_{S}(x) - f_{T}(x)\|], \mathcal{E}_{D_{T}}[\|f_{S}(x) - f_{T}(x)\|] \}
\]

(1)

Our previously proposed DGA-DA provides a unified framework which jointly optimizes term 2 in Eq.(1) as DDC-DA methods when aligning data distributions, and term 3 in Eq.(1) as DGSA-DA approaches when performing label inference through the underlying data geometric structure. It further introduces a repulsive force (RF) term using both source and target domain data when seeking the latent feature space and makes the proposed DA method discriminative. In this paper, we go one step further and propose a novel DA method, namely Discriminative Noise Robust Sparse Orthogonal Label Regression-based Domain Adaptation (DOLL-DA), which optimizes at the same time the three terms of the right-hand of Eq.(1), including in particular the first term on classification error in Eq.(1), when seeking a discriminative latent feature space.

Specifically, the proposed DOLL-DA derives from an integrated DA model which: 1) searches a shared feature subspace where the source and target data distributions are discriminatively aligned using an improved repulsive force (RF) term added to the MMD constraints, thereby optimizing the second term in Eq.(1) and indirectly improving its first term; 2) makes use an embedding trick to immerge data labels into the shared feature space and projects each data sample within the vicinity of its label vector orthogonal to other ones, thereby further regularizing the improved MMD constraints and avoiding potential contradictions among sub-domains when repulsive force is applied in the search of the shared feature subspace; 3) linearly regresses data labels in the shared feature subspace, thereby further explicitly optimizes the first term of the right-hand in Eq.(1). Moreover, data outliers are accounted for and the property of sparsity in label regression is introduced to circumvent negative transfer and over-fitting, leading to a novel Noise Robust Sparse Orthogonal Label Regression (NRS_OLR) term in our DA model; 4) leverages the true labels available in the source domain and ensures a label consistency between the source and target domain through an iterative integrated linear label regression, thereby minimizing the third term of Eq.(1); Fig.1 depicts the general framework of the proposed DOLL-DA method.

To sum up, the contributions of this paper are as follows:

- Orthogonal label subspace is proposed through a label embedding trick to further regularize the improved RF-based MMD constraints using a novel Orthogonal Label Regression (ORL) constraint, thereby circumventing potential conflicts which could arise when optimizing the improved RF-based MMD constraints and further enhancing the discriminative power of the proposed DA method.
- The hypothesis for classifying the target domain data is learned simultaneously through a single feature projection matrix \( A \) when aligning discriminatively the source and target domain data in the shared feature subspace. Furthermore, a property of sparsity is introduced through a \( l_{2,1} \)-norm constraint on \( A \) when regressing data labels and data outliers are also accounted for within the model, leading to a Noise Robust Sparse Orthogonal Label Regression (NRS_OLR) term to ensure the proposed DA model to avoid negative transfer as well as overfitting.
- A novel generalized power iteration method is introduced to solve the optimization problem of the full integrated DA model, resulting in the proposed DOLL-DA. Furthermore, we perform time complexity analysis and also derive three additional DA methods based on three partial models from the full integrated DA in order to highlight the individual contribution of RF and NRS_OLR term, respectively, as well as their added value when they are jointly optimized.
- We perform extensive experiments on 49 image classification DA tasks using 8 popular DA benchmarks and demonstrate the effectiveness of the proposed DOLL-DA which consistently outperforms thirty state-of-the-art DA algorithms with a margin which can reach 17 points. Moreover, we also carry out in-depth analysis of the proposed DA method, in particular w.r.t. their hyperparameters, convergence speed, the choice of the base classifier, random label initialization and impact of the quantity of target domain data for performance stability.

The article is organized as follows. Sect.II discusses the related work. Sect.III presents the method. Sect.IV benchmarks the proposed DA method and provides in-depth analysis. Sect.V draws the conclusion.

II. RELATED WORK

Last years have seen DA techniques applied to multiple computer vision applications. State-of-the-art has so far featured two main research streams: 1) Shallow DA; 2) Deep DA. They are overviewed in Sect.II-A and sect.II-B, respectively, and discussed in comparison with the proposed DOLL-DA in sect.II-C.

A. Shallow Domain Adaptation

1) Feature-based DA: The rationale of the feature-based domain adaptation is to assume a shared latent feature space between the source and target domain which is searched in narrowing the existing distribution discrepancies across the domains. A popular strategy for searching such a shared latent feature space is to embrace the dimensionality reduction and
propose to explicitly minimize some predefined distance measure to reduce the mismatch between source and target in terms of marginal distribution [52] [42], or conditional distribution [33]. These methods can be further distinguished based on whether they incorporate some form of data discriminativeness or not in the search of such a shared latent feature space.

**Nondiscriminative distribution alignment (NDA):** NDA strategies propose to align the marginal and conditional distributions across the source and target domains in reducing different data distribution distance measurements, e.g., Bregman Divergence [52], Wasserstein distance [6], [7], Maximum Mean Discrepancy (MMD) to explicitly shrink the cross-domain divergence of marginal data distributions [42], and both the marginal and conditional data distributions [33]. The obvious drawback of NDA is that it ignores the discriminative knowledge among different labeled source sub-domains, thereby increasing the burden of the required classifier.

**Discriminative distribution alignment (DDA):** DDA methods improve NDA ones by explicitly leveraging the discriminative information in source data labeled into different sub-domains. ILS[20] learns a discriminative latent space using Mahalanobis metric and makes use of Riemannian optimization strategy to match statistical properties across different domains. [36] adapts linear discriminant analysis (LDA) and leverages the discriminative information from the target domain to estimate the common feature space. OBTL,[25] proposes bayesian transfer learning based domain adaptation, which explicitly discusses the relatedness across different sub-domains. SCA[15] achieves discriminativeness in optimizing the interplay of the between and within-class scatters. Our proposed DGA-DA[39] also introduces a specific repulsive force term to capture the data discriminativeness. However, our DGA-DA further cares about the underlying data manifold structure when performing label inference. However, despite improvement over NDA methods, DDA methods rely upon temporary and unreliable pseudo labels of target domain data to capture the repulsive force in the target domain, and thereby can mislead the search of an optimized shared discriminative latent space.

2) **Subspace alignment-based DA:** In line with [12], an increasing number of DA methods, e.g., [40], [51], [63], [54], [9], emphasize the importance of aligning the underlying data subspace and 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 [51], or interleaved by source samples [63], thereby aligning the geometric structures of the underlying data manifolds. A few recent DA methods, e.g., RSA-CDDA[40], JGSA[64], further propose unified frameworks to reduce the shift between domains both statistically and geometrically. HCA[31] improves JGSA using a homologous constraint on the two transformations for the source and target domains, respectively, to make the transformed domains related and hence alleviate negative domain adaptation.

However, in light of the upper error bound as defined in Eq.(1), we can see that subspace alignment based DA methods account for the underlying data geometric structure and expect but without theoretical guarantee the alignment of discriminative data distributions. Our proposed DOLL-DA improves subspace alignment-based DA by jointly aligning the
data distributions discriminatively and enhancing the hidden manifold structure through label regression to regress different sub-domains w.r.t its orthogonal label subspace.

B. Deep Domain Adaptation

Recently, DA has been intensively investigated under the paradigm of deep learning (DL), and has featured the following two main approaches.

1) Statistic matching-based DA: These methods aim to reduce the divergence across domains using statistic measurements incorporated into DL frameworks. DAN[32] reduces the marginal distribution divergence in incorporating the multi-kernel MMD loss on the fully connected layers of AlexNet. JAN[35] improves DAN by jointly decreasing the divergence of both the marginal and conditional distributions. D-CORAL[55] further introduces the second-order statistics into the AlexNet[27] framework for more effective DA strategy.

2) Adversarial loss-based DA: These methods make use of GAN[17] and propose to align data distributions across domains in making sample features indistinguishable w.r.t the domain labels through an adversarial loss on a domain classifier [14], [57], [46]. DANN[14] and ADDA[57] learn a domain-invariant feature subspace in reducing the marginal distribution divergence. MADA [46] additionally make use of multiple domain discriminators, thereby aligning conditional data distributions. Different from the previous approaches, DSN[4] achieves domain-invariant representations in explicitly separating the similarities and dissimilarities in the source and target domains. MADAN[65] explores knowledge from different multi-source domains to fulfill DA tasks. CyCADA[22] addresses the distribution divergence using a bi-directional GAN based training framework.

The main advantage of these DL based DA methods is that they jointly shrink the divergence of data distributions across domains and achieve a discriminative feature representation of data through a single unified end-to-end learning framework. However, they also present the drawback that the discriminative force is merely extracted from the labelled source domain while it could rely upon simultaneously the source and target domains in leveraging the underlying data geometric structures. Furthermore, these DL based DA approaches mostly work as a black-box and suffer from interpretability, thereby falling short to provide deep insights for further improving DA methods.

C. Discussion

Fig.2 compares our proposed DOLL-DA with our previously proposed DA method, DGA-DA[39] as well as MEDA[62], and highlights their similarities and differences according to the following 6 properties:

- **Dis=(Discriminativeness):** both DGA-DA and DOLL-DA introduce a (RF) term to achieve discriminativeness across domains, while MEDA ignores this merit. DGA-DA takes into account RF term across domain, while the proposed DOLL-DA improves it in extending it with a RF term within the source domain.

- **Reg=(Regularization):** Different from DGA-DA and MEDA, DOLL-DA proposes orthogonal label regression which further regularizes the effectiveness of the improved RF-based MMD constraints, thereby circumventing potential contradictions when reinforcing these MMD constraints.

- **Cons=(Constraint):** Both DGA-DA and DOLL-DA are optimized using the constraint $A^T X H X^T A = I$, which removes an arbitrary scaling factor in the embedding and prevents the optimization from collapsing into a subspace of dimension less than the required dimensions. MEDA does not have such a constraint to obtain an analytical solution.

- **Joint=(Joint optimization):** Both MEDA and DOLL-DA achieve data distribution alignment and classifier optimization through a unified joint optimization model, whereas the optimization in DGA-DA for data distribution alignment and classifier optimization is carried out separately.

- **Error:** DOLL-DA designs an error tolerated subspace to care about the noise in data, therefore decreasing the risk of potential negative transfer.

- **Manifold:** both MEDA and DGA-DA capture the underlying data manifold structures for label inference. While effective, they also suffer from the computational burden due to the singular value decomposition. DOLL-DA also cares about data manifold structures, but through computation effective iterative integrated linear label regression.

### III. The Proposed Method

Sect.III-A defines the notations and states the DA problem. Sect.III-B formulates our DA model while Sect.III-C presents the generalized power iteration method to solve the proposed DA model and derives the algorithm of DOLL-DA. Sect.III-D extends DOLL-DA to non-linear problem through kernel mapping. Sect.III-E performs time complexity analysis of the proposed DOLL-DA.

#### A. Notations and Problem Statement

Matrices are written as boldface uppercase letters. Vectors are written as boldface lowercase letters. For matrix $M = (m_{ij})$, its $i$-th row is denoted as $m^i$, and its $j$-th column is denoted by $m_j$. We define the Frobenius norm $\|\cdot\|_F$ and $l_{2,1}$ norm as: $\|M\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{l} m_{ij}^2}$ and $\|M\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{l} m_{ij}^2}$. A domain $D$ is defined as an l-dimensional feature space $\chi$ and a marginal probability distribution $P(x)$, i.e., $D = \{\chi, P(x)\}$ 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 $f(x) = Q(y | x)$ can be interpreted as the class conditional probability distribution for each input sample $x$.

In unsupervised domain adaptation, we are given a source domain $D_S = \{x_1, y_1, i = 1, \ldots, n_i\}$ with $n_i$ labeled samples $X_S = \{x_1, \ldots, y_n\}_c \in \mathbb{R}$, which are associated with their class labels $Y_S = \{y_1, \ldots, y_n\}_c \in \mathbb{R}$, and an unlabeled target domain $D_T = \{x_1, y_1, i = 1, \ldots, n_i\}_c$ whose labels $Y_T = \{y_{n_i+1}, \ldots, y_{n+n_i}\}_c \in \mathbb{R}$ are unknown. Here, $y_i \in \mathbb{R}^c (0 \leq i \leq n_i)$ is a one-vs-all label hot vector in which $y_i^c = 1$ if $i \in T$ belongs to the $c$-th class, and 0 otherwise. We define the data matrix $X = [X_S, X_T] \in \mathbb{R}^{l \times n}$ ($l = \text{feature dimension}; n = n_s + n_t$) in packing both the source and target data. The source domain $D_S$ and target domain $D_T$ are assumed to be different, i.e., $\chi_S = \chi_T$, $\chi_S = \chi_T$, $\mathcal{P}(\chi_S) \neq \mathcal{P}(\chi_T)$, $Q(\chi_S | \chi_S) \neq Q(\chi_T | \chi_T)$. We also define the notion of sub-domain, i.e., class, denoted as $D_S^c$, representing the set of samples in $D_S$ with the class label $c$. It is worth noting that, the definition of sub-domains in the target domain, namely $D_T^c$, requires a base classifier, e.g., Nearest Neighbor (NN), to attribute pseudo labels for samples in $D_T$.

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 into Reproducing Kernel Hilbert Space (RKHS). Given two distributions $P$ and $Q$, the MMD between $P$ and $Q$ is defined as:

$$\text{Dist}(P, Q) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(p_i) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(q_i) \right\|_H$$

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

The aim of the proposed DOLL-DA is to search jointly a transformation matrix $A \in \mathbb{R}^{l \times k}$ projecting discriminatively both the source and target domain data of dimension $l$ into a latent shared orthogonal feature subspace of dimension $k$ as well as a label regressor while minimizing simultaneously the three terms of the upper bound error in Eq.(1).

**B. Formulation**

Our final model DOLL-DA (sect.III-B5) starts from JDA (sect.III-B1), which is improved in Sect.III-B2 and Sect.III-B3 for discriminative data distribution alignment (DDA) by leveraging the discriminative knowledge from the source and target domains. Fig. 3 summarizes these steps which aim to minimize $d_\mathcal{H}(D_S, D_T)$ (term 2 of Eq.(1)). Sect.III-B4 further introduces an orthogonal label regressor using an embedding trick and accounts for noisy data as well as sparsity in label regression to derive Noise Robust Sparse Orthogonal Label Regression (NRS-OLR). Sect.III-B5 integrates DDA and NRS-OLR to achieve our final model and thereby optimizes at the same time the three terms of the right-hand of Eq.(1).

1) Matching Marginal and Conditional Distributions: As shown in Fig.3.a and Fig.3.b, our model starts from JDA, which makes use of MMD in RKHS to measure the distances between the expectations of the source domain/sub-domain and target domain/sub-domain. Specifically, 1) The empirical distances of the source and target domains is defined as $\text{Dist}^s$; 2) The conditional distance $\text{Dist}^c$ is defined as the sum of the empirical distances between sub-domains in $D_S$ and $D_T$ with a same label; 3) $\text{Dist}_{ClO}$ is defined as the sum of $\text{Dist}^s$ and $\text{Dist}^c$.

$$\text{Dist}_{ClO} = \text{Dist}^s(D_S, D_T) + \text{Dist}^c(D_S^c, D_T^c)$$

2) Across domain Repulsive force term: As shown in Fig.3.(a,b), sect.III-B1 merely cares about shrinking the $\text{MMD}$ and $\text{Dist}^c$ across source domain/sub-domain and target domain/sub-domain. Specifically, 1) 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 into Reproducing Kernel Hilbert Space (RKHS) (RKHS). Given two distributions $P$ and $Q$, the MMD between $P$ and $Q$ is defined as:

$$\text{Dist}(P, Q) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(p_i) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(q_i) \right\|_H$$

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

The aim of the proposed DOLL-DA is to search jointly a transformation matrix $A \in \mathbb{R}^{l \times k}$ projecting discriminatively both the source and target domain data of dimension $l$ into a latent shared orthogonal feature subspace of dimension $k$ as well as a label regressor while minimizing simultaneously the three terms of the upper bound error in Eq.(1).

**B. Formulation**

Our final model DOLL-DA (sect.III-B5) starts from JDA (sect.III-B1), which is improved in Sect.III-B2 and Sect.III-B3 for discriminative data distribution alignment (DDA) by leveraging the discriminative knowledge from the source and target domains. Fig. 3 summarizes these steps which aim to minimize $d_\mathcal{H}(D_S, D_T)$ (term 2 of Eq.(1)). Sect.III-B4 further introduces an orthogonal label regressor using an embedding trick and accounts for noisy data as well as sparsity in label regression to derive Noise Robust Sparse Orthogonal Label Regression (NRS-OLR). Sect.III-B5 integrates DDA and NRS-OLR to achieve our final model and thereby optimizes at the same time the three terms of the right-hand of Eq.(1).
Optimization Strategies

Visualization

Discrepancy Measurements

Marginal distribution approximation

Conditional distribution approximation

Minimization Problem

Discrepancy Measurements

Marginal distribution approximation

Conditional distribution approximation

Fig. 3: (a) MMD constraints; (b) Complex relationships; (c) 2-classes DA with MMD; (d) N-classes DA with MMD.

Fig. 4: Illustration of MMD constraints and orthogonal projection based label consistent regression.

4) Noise Robust Sparse Orthogonal Label Regression: The improved repulsive force term formulated in Sect.III-B2 and Sect.III-B3 aims to increase the between-class distances, across domain through Eq.(4) and within the source domain through Eq.(5).
Eq.(5), and enable discriminative alignment of marginal and conditional data distributions (eq.(3)) in sect.III-B1), but does not seek to decrease the intra-class distances. As a result, situations as shown in Fig.4.(a,b,c), where the instances of a class are pushed away from other class instances but don’t get closer each other within the class, could happen. More importantly, while eq.(4) and eq.(5) make it possible to decrease the classification errors of a hypothesis on the source domain, the first term of the error bound in eq.(1) is not explicitly optimized. Here, we propose to solve the previous two issues through orthogonal label regression as shown in Fig.4.(d, e, f) where the instances of each class (sub-domain) across domain get close to their one-vs-all hot label vector which is orthogonal to those of other classes.

To this end, we introduce a novel orthogonal label regression (OLR) constraint $\Phi(A, Y_S, X_S)$, where $A$ is the transformation matrix projecting both the source and target data onto a latent shared feature subspace of dimension $k$. Specifically, we first introduce an embedding trick which consists of immersing a $C$-dimensional one-vs-all hot label vector into the $k$-dimensional latent shared feature subspace simply by adding $(k - C + 1)$ times $0, e.g., a 3$-dimensional one-vs-all label vector $(0, 1, 0)^T$ is represented by its corresponding one-vs-all hot vector $(0, 1, 0, 0, 0)^T$ in the $5$-dimensional feature subspace using the embedding trick. We can then perform least square regression (LSR): \( \min \| X^T A - Y \|_F^2 \) st. \( Y \geq 0 \), \( Y 1 = 1 \), with $Y$ the class label matrix as defined in sect.III-A and extended into a $n \times k$ matrix by embedding each one-vs-all hot label vector into a $k$-dimensional one using the embedding trick. Minimization of LSR thus simply requires that each labeled sample be projected within the vicinity of its corresponding label hot vector in the $k$-dimensional feature space.

It is worth noting that the proposed OLR enjoys the following three properties:

- **Orthogonality**, i.e., \( (Y_{i=1..C} \cdot Y_{i=\neq C}) = 0 \) with $\cdot$ denoting the dot product. This constraint simply expresses that one-vs-all hot label vectors in the shared latent feature subspace are orthogonal each other. As a result, projecting each data sample into the vicinity of its corresponding label vector keeps the samples of each sub-domain far away from those of other sub-domains and thereby improving data discriminativeness and optimizing term.1 of Eq.(1);

- **Label Embedding Constraint**, i.e., $Q(Y_{i=1..C} | X_{S \cup X_T}) = 0$. This property simply denotes the fact that we have made use of the embedding trick for immersing each $C$ dimensional one-vs-all hot label vector into the $k$ dimensional shared feature space and there are no label vectors for classes ranging from $C + 1$ to $k$;

- **Sharing of the feature projection and label regression matrix** through $A$. Thanks to the label embedding constraint, the projection matrix $A$ through eq.(3), eq.(4) and eq.(5) for the search of a shared latent feature space aligning discriminatively marginal and conditional data distributions between the source and target domain can be shared with the one used for the orthogonal label regression (OLR) constraint, thereby jointly optimizing term.1 and term.2 of Eq.(1) within a single unified feature and label subspace. Furthermore, as shown in Fig.4.(d), (e)), data of each sub-domain, i.e. data with a same label, from the source and target domain, are projected within the vicinity of their corresponding one-vs-all hot label vector, thereby also decreasing Term.3 of Eq.(1), i.e., the errors of their respective labelling functions on the source and target domain.

However, data from the source and target domain can be noisy. We account for data noise through an error matrix $E$ and the OLR constraint can therefore be reformulated as: \( \min \| X^T A - Y + E \|_F^2 \) st. \( Y \geq 0 \), \( Y 1 = 1 \). The error matrix $E$ makes possible a certain tolerance of errors when projecting data into the vicinity of its corresponding label vector in the latent shared feature subspace, thereby enabling to account for outliers and alleviating the influence of negative transfer. Additionally, given the fact that, in real-life applications, e.g. visual object recognition, data of a given class generally lie within a manifold of much lower dimension in comparison with the original data space, e.g., pixel number of images, we further introduce a $l_2,1$-norm constraint so as to fulfill the property that the class label of a data sample should be regressed from a sparse combination of features in the latent shared feature subspace. This constraint introduces a regularization term on $A$ for discriminative subspace projection, which also optimizes Term.3 of Eq.(1). Putting all these together, the initial Orthogonal Label Regression (OLR) constraint becomes Noise Robust Sparse Orthogonal Label Regression (NRS-OLR) and is finally formulated as:

\[
\min \| X^T A - Y + 1e^T \|_F^2 + \beta \| A \|_F^2 \text{ st. } Y \geq 0, Y 1 = 1, (Y_{i=1..C} \cdot Y_{i=\neq C}) = 0
\]

5) The final model: By integrating all the properties introduced in sect.III-B1 through sect.III-B4, we obtain our final DA model, formulated as Eq.(19)

\[
A.s.t. Y_{i=0..C} Y_{i=1} = 0 \\
(\text{tr}(A^T X M^+ X^T A) + \alpha \| A \|_F^2 + \beta \| A \|_F^2) + \| X^T A - Y + 1e^T \|_F^2 + \| X^T A - Y + 1e^T \|_F^2
\]

\[
\text{st. } M^+ = M_0 + \sum_{c=1}^C (M_c - M_{REF}), Y \geq 0, Y 1 = 1, A^T X H X^T A = I, (Y_{i=1..C} \cdot Y_{i=\neq C}) = 0
\]

where $M_{REF} = M_{S \rightarrow T} + M_{T \rightarrow S} + M_{S \rightarrow S}$ is the improved overall repulsive force constraint matrix, $H = I - \frac{1}{n}$ is the centering matrix, the constraint $Y \geq 0$ and $Y 1 = 1$ simply expresses the fact that each data sample has a label vector whose class probability sums to 1, whereas the constraint $A^T X H X^T A = I$ derives from Principal Component Analysis (PCA) to preserve the intrinsic data covariance of both domains and avoid the trivial solution for $A$.

Through iterative optimization of Eq.(19), our DA method searches jointly a well regularized discriminative latent feature subspace shared between the source and target domain and a noise robust sparse label regression model, thereby optimizing at the same time the three terms of the right-hand of Eq.(1).

C. Solving the model

Eq.(19) is not convex, therefore we propose an effective method which optimizes the key variables, e.g., $A, M, Y$, in
a coordinate descent manner. Main steps for solving Eq.(19)
are as follows. All the key steps have closed form solution:

**Step.1** (Initialization of \( M^* \)) \( M^* \) can be initialized by calculating \( M_0 \) since there is no labels or pseudo labels on the target domain initially. We obtain \( M^* = M_0 \) where \( M_0 \) is the MMD matrix as defined in Eq.(3).

**Step.2** (Initialization of \( A \)) Similar to JDA, \( A \) can be initialized to reduce the marginal and conditional distributions between \( P(X_S) \) and \( P(X_T) \) through an adaptive feature subspace via the Rayleigh quotient algorithm, solving Eq.(8):

\[
(X(M_0 + \sum_{c=1}^C M_c)X^T + \alpha I) A = XHX^T A \Phi
\]

where \( X = X_S \cup X_T \) and \( \Phi = diag(\phi_1, ..., \phi_k) \in R^{k \times k} \) are Lagrange multipliers, \( \sum_{c=1}^C M_c \) is obtained through labeled source domain data \( A'X_S \) and pseudo labels inferred on the target domain data \( A'X_T \). \( A \) is then initialized as the k smallest eigenvectors of Eq.(8), with \( k \) defining the dimension of the latent shared feature subspace between the source and target domain.

**Step.3** (Update of \( e \)) \( e \) is updated in solving Eq.(19) with other variables held fixed. To update \( e \), one should solve Eq.(9)

\[
e' = \arg \min_e ||X^T A + e_1 e - Y||^2_F
\]

In setting to 0 the partial derivative of Eq.(9) with respect to \( e \), we achieve the optimal solution of \( e \) as

\[
e = \frac{1}{n}(Y^T 1 - A^T X 1)
\]

**Step.4** (Update of \( A \)) \( A \) is updated by solving the optimization problem in Eq.(19) with other variables held fixed. To ensure that Eq.(19) is differentiable, we regularize \( ||A||^2_{2,1} \) as

\[(\sum_{j=1}^k \sqrt{||a_j||^2_2 + \varepsilon})\] to avoid \( ||A||^2_{2,1} = 0 \). As a result, Eq.(19) becomes Eq.(11)

\[
A' = \min_{A, A^T X X^T A = I} \{ \alpha ||A||^2_F + \beta (\sum_{j=1}^k \sqrt{||a_j||^2_2 + \varepsilon}) \}
\]

\( \varepsilon \) is infinitely close to zero, making Eq.(11) closely equivalent to Eq.(19). Solving directly Eq.(11) is non-trivial, we introduce a new variable \( G \in R^{k \times k} \) which is a diagonal matrix with

\[
g_{jj} = (\sum_{j=1}^k \sqrt{||a_j||^2_2 + \varepsilon}) / \sqrt{||a_j||^2_2 + \varepsilon}.
\]

\( G \) and \( A \) can be optimized iteratively. With \( G \) held fixed and \( e \) computed as in Eq.(10), we can reformulate Eq.(11) as Eq.(12)

\[
A' = \min_{A, A^T X X^T A = I} \{ \alpha ||A||^2_F + \beta (\sum_{j=1}^k \sqrt{||a_j||^2_2 + \varepsilon}) \}
\]

Eq.(12) is a least squares problem on the Stiefel manifold, which is a non-convex optimization problem. Therefore, it cannot be directly solved via the Lagrangian method or an analytical solution. Inspired by previous research on solving quadratic problem on the Stiefel manifold[13], [41], [19], we propose a novel generalized power iteration (GPI) method to optimize the projection matrix \( A \) that rotates the factor matrix to best fit the hypothesis subspace.

**Step.i** We propose Cholesky factorization of \( H \), which aims to obtain a lower triangular matrix \( h \), so that \( hh^T = H \).

**Step.ii** Eq.(12) is reformulated as:

\[
\arg \min_{A, A^T X X^T A = I} \{ \alpha ||A||^2_F + (X(M^*)X^T)A) - 2 \beta (A^T X H^T Y) \}
\]

We set:

\[
W^T = A^T X h
\]

\( C = h_h^T H Y \)

\( B = (X^T) + \beta (X^T X) \)

Using Eq.(14), Eq.(13) can be written for short as:

\[
\arg \min_{W, W^T} \{ tr(W^T B W - 2 W^T C) \}
\]

where \( B \) is a symmetric matrix.

**Step.iii** Initialize \( W^T = A^T X h \) to satisfy \( W^T W = I \), and set \( B = \mu I - B \), which ensures \( B \) is a positive definite matrix.

**Step.iv** Set \( Z = 2 B^T W + 2 C \). Then, optimize \( USV^T = Z \) via singular value decomposition method on \( Z \).

**Step.v** Update \( W^T = U V^T \) and \( A^T = W^T (X h)^{-1} \).

Eventually, the final optimization of Eq.(11) is \( A' = (W^T (X h)^{-1})^T \). Algorithm 1 details the whole process to update \( A' \).

**Algorithm 1:** Power iteration method for solving Eq.(11)

**Input:** Data \( X \). Source domain label \( Y_S \), MMD matrix \( M^* \), fixed matrix \( G, H \), regularization parameters \( \beta \) and \( \alpha \)

1: Initialize \( W \) and \( B' \) as introduced in **Step.iii**;
2: Update \( Z, A^T, W^T \);
3: if Non convergence then
4: (i) Do **Step.iv**;
5: (ii) Do **Step.v**;
6: else
7: break;
8: Update \( A' \) via solving \( A' = (W (X h)^{-1})^T \)

**Output:** \( A \leftarrow A' \)

**Step.5** (Update of \( Y \)) The label matrix \( Y \) contains two parts: true labels \( Y_S = \{y_1, ..., y_m\}^T \in R^{n \times C} \), and pseudo labels \( Y_T = \{y_{n+1}, ..., y_{n+n}\}^T \in R^{n \times C} \). Our aim is to iteratively refine the latter ones. Given fixed \( A \) and \( M^* \), each \( y_i \in Y_T \) can be updated by solving the following problem:

\[
y'_i = \min_{y_i \in R^C} ||X^T A - y_i + e||^2_F
\]

Using Lagrangian multipliers method, the final optimal solution of \( y_i \) is

\[
y'_i = (A^T X_i + e + \theta)
\]

where \( \theta \) is coefficient of Lagrangian constraint \( y_i^T 1 - 1 = 0 \), which can be obtained by solving \( y'_i 1 = 1 \).

**Step.6** (Update of \( M^* \)) With the labeled source domain data \( A'X_S \) and the labels inferred on the target domain data \( A'X_T \) as in **Step.5**, we can update \( M^* \) as

\[
M^* = M_0 + \sum_{c=1}^C (M_c - M_{REP})
\]

where \( M_c \) and \( M_{REP} \) are defined in Eq.(19).

The complete learning algorithm is summarized in Algorithm 2 - DOLL-DA.
**Algorithm 2**: Discriminative Label Consistent Domain Adaptation (DOLL-DA)

**Input**: Data $X$, Source domain label $Y_s$, subspace dimension $k$, iterations $T$, regularization parameters $\beta$ and $\alpha$

1. Initialize $M^* = M_0$ as defined in Eq.(3);
2. Initialize $A$ by solving Eq.(8); ($t := 0$)
3. while $\sim \text{isempty}(X,Y_s)$ and $t < T$ do
   3.1: Update $M^*$ by solving Eq.(18)
   3.2: Update $e$ by solving Eq.(10)
   3.3: Update $A$; ($t_1 := 0$.)
   3.4: if $t_1 < T$ then
      3.4.1: Update $G$ as an identity matrix;
      3.4.2: Update $A$ by solving Algorithm1;
      3.4.3: Update $G$ by calculating
      \[ g_{jj} = \left( \sum_{i=1}^{k} \| \alpha_i^T \|_2^2 + \varepsilon \right) \div \left( \sqrt{\| \alpha_i^T \|_2^2 + \varepsilon} \right); \]
      3.4.4: ($t_1 = t_1 + 1$);
   3.5: else
      \[ \text{break}; \]
   3.6: Update $Y$ by solving Eq.(17)
   3.7: Update pseudo target labels
      \[ Y^T_{(t)} = Y \left[; (n_s + 1) : (n_s + n_t) \right]; \]
   3.8: $t = (t + 1)$;
4. Output: $A$, $Z = A^T X$, $Y$

**D. Kernelization**

The proposed DOLL-DA method is extended to nonlinear problems in a Reproducing Kernel Hilbert Space via the kernel mapping $\phi : x \rightarrow \phi(x)$, or $\phi(X) : [\phi(x_1), ..., \phi(x_n)]$, and the kernel matrix $K = \phi(X)^T \phi(X) \in R^{n \times n}$. We utilize the representer theorem to formulate the Kernel DOLL-DA as:

\[
\min_{A,0 \leq Y^T \geq 0,Y^T = 1} (tr(A^T K M^* K^T A) + \alpha \| A \|_F^2 + \beta \| A \|_2^2) + \| K^T A + 1 \varepsilon^T - Y \|_F^2
\]

\[
st.M^* = M_0 + \sum_{t=1}^{T} (M_o) - M_{REP}, \ Y \geq 0, \ Y1 = 1,
A^T K H K^T A = 1, (Y_{inc} \bullet Y_{inc}) = 0 \tag{19}
\]

**E. Time Complexity Analysis**

Given $n$ the number of data samples including both the source and target domain and $l$ the feature dimension, we denote by $t$, $t_1$ the number of iterations with $t, t_1 \sim \min(l,n)$. The major computational burden of the proposed Algorithm.2 lies in Step 2, 3 and 5 as sketched in Sect.III-C. In Step 2, the singular value decomposition(SVD) is computed on a $n \times n$ matrix, its computational complexity is $O(n^3)$. Step 3. constructs the $M_{case}$ matrix, whose computational complexity is $O(4CN^2)$ with $C$ the number of classes, i.e., the class cardinality. Step 4.(ii) makes use of SVD to solve optimization on a $n \times k$ size matrix as introduced in step.(iv) of Algorithm.1, its computational complexity is thus $O(n^2k + nk^2 + k^3)$. In Step 5 and Step.6, $O(nk)$ operations are required for all other lines. Therefore, the overall computational complexity of the proposed Algorithm - DOLL-DA is $O(n^3 + 4tCn^2 + tnk + tt_1(n^2k + nk^2 + k^3))$.

**IV. Experiments**

**A. Benchmarks and Features**

As illustrated in Fig.5, USPS[23]+MINIST[28], COIL20[33], PIE[33], office+Caltech[33], Office-Home[60] and SVHN-MNIST[4] 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[15], [8], [40], [4], [30] do. We construct 49 datasets for different image classification tasks.

**Office+Caltech** consists of 2533 images of 10 categories (8 to 151 images per category per domain)[15]. 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[16] 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)[11] are deep features computed as in AELM[59] which makes use of VLSMatConvNet library with different pretrained CNN models, including in particular the Caffe implementation of AlexNet[27] 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. The arrow $\rightarrow$ is proposed to denote the direction from “source” to “target”. For example, “W $\rightarrow$ D” means the Webcam image dataset is consdered as the labeled source domain whereas the DSLR image dataset the unlabeled target domain.

**USPS+MNIST** shares 10 common digit categories from two subsets, namely USPS and MNIST, but with very different data distributions (see Fig.5). We construct a first DA task USPS $vs$ MNIST by randomly sampling first 1,800 images in USPS to form the source data, and then 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 $16 \times 16$, and represent each one by a feature vector encoding the gray-scale pixel values. We also extract deep feature from softmax layer[47] of LeNet[28] architecture, leading to a 10 dimensional feature. Thus the source and target domain 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.5). 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 $32 \times 32$ pixels and 256 gray levels per pixel. In this experiment, we partition the dataset into two subsets,
namely COIL 1 and COIL 2[63], COIL 1 contains all images taken within the directions in $[0^\circ, 85^\circ] \cup [180^\circ, 265^\circ]$ (quadrants 1 and 3), resulting in 720 images. COIL 2 contains all images taken in the directions within $[90^\circ, 175^\circ] \cup [270^\circ, 355^\circ]$ (quadrants 2 and 4) and thus the number of images is also 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 $\rightarrow$ COIL2 and COIL2 $\rightarrow$ COIL1.

PIE face database consists of 68 subjects with each under 21 various illumination conditions[8], [33]. We adopt five pose subsets: C05, C07, C09, C27, C29, which provide a rich basis for domain adaptation, that is, we can choose one pose as the source and any remaining 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 $32 \times 32$ 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.

Office-Home dataset as shown in Fig.5 is a novel DA dataset recently introduced in [60]. This dataset contains 4 domains. Each domain contains 65 categories. This dataset is used in a similar manner as the Office+Caltech dataset. From the 4 domains, i.e., the Art (Ar), Clipart (Cl), Product (Pr) and Real-World (Rw), we generate 12 DA tasks, namely namely $Ar \rightarrow Cl, . . . Pr \rightarrow Rw$, respectively. DeCAF6 features are extracted to evaluate the performance of the proposed DA algorithms on this dataset.

SVHN-MNIST contains the MNIST dataset as introduced in USPS-MNIST but also the Street View House Numbers (SVHN) which is a collection of house numbers collected from Google street view images (see Fig.5). SVHN is quite distinct from the dataset of handwriting digits, i.e., digits in MNIST. Moreover, all the 2 domains are quite large, each having at least 60k samples over 10 classes. We propose to make use of the LeNet architecture [28] and domain classifier as introduced in [47] to extract features for our DA tasks.

B. Baseline Methods

The proposed DOLL-DA method is compared with thirty-two methods of the literature, including deep learning-based approaches for unsupervised domain adaption. They are:

- Shallow methods: (1) 1-Nearest Neighbor Classifier(NN); (2) Principal Component Analysis (PCA); (3) GFK [16]; (4) TCA [42]; (5) TSL [52]; (6) JDA [33]; (7) ELM [59]; (8) AELM [59]; (9) SA [12]; (10) mSDA [5]; (11) TJM [34]; (12) RTML [8]; (13) SCA [15]; (14) CDML [61]; (15) LTSL [51]; (16) LRSR [63]; (17) KPCA [49]; (18) JGSA [64]; (19) CORAL [54]; (20) RVDLR [24]; (21) LPJT [29]; (22) DGA-DA[39].

- Deep methods: (23) AlexNet [27]; (24) DAH [60]; (25) DANN [14]; (26) ADDA [57]; (27) LTRU [50]; (28) ATU [48]; (29) BSWD [47]; (30) DSN [4]; (31) DDC [58]; (32) DAN [32].

Direct comparison of the proposed DOLL-DA using shallow features against these DL-based DA approaches could be unfair. However, in order to give an idea on the performance gap between shallow and deep DA methods, we still compare their results with those of our shallow DOLL-DA. For the purpose of fair comparison, we follow the experiment settings of DGA-DA, JGSA and BSWD, and apply DeCAF6 as the input features for some methods to be evaluated. Whenever possible, the reported performance scores of the thirty-two methods of the literature are directly collected from their original papers or previous research [57], [59], [29], [15], [47], [64], [39]. 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[39], [33], [63], we also evaluate the proposed DOLL-DA by empirically searching in the parameter space for the optimal settings. Specifically, the proposed DOLL-DA method has three hyper-parameters, i.e., the subspace dimension $k$, regularization parameters $\beta$ and $\alpha$. In our experiments, we set $k=300$ and $1) \beta=0.1$, and $\alpha=1$ for USPS, MNIST · COIL20 and PIE, 2) $\beta=1$, $\alpha=1$ for Office+Caltech, SVHN-MNIST and Office-Home.

In our experiment, accuracy on the test dataset as defined by Eq.(20) is the performance measurement. It is widely used in literature, e.g.,[32], [40], [33], [63], etc.

\[
\text{Accuracy} = \frac{|\{x:x\in D_T \land \hat{y}(x) = y(x)\}|}{|\{x:x\in D_T\}|} 
\]

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

The core model of the proposed DOLL-DA method is built on JDA, but adds up two optimization terms, namely discriminative data distribution alignment (DDA) term as defined in Eq.(4, 5), and Noise Robust Sparse Orthogonal
Label Regression (NRS\textsubscript{OLR}) term as defined in Eq.\,(6). For further insight into the proposed DOLL-DA and the rational w.r.t.\ the DDA and NRS\textsubscript{OLR} term, respectively, we derive from Eq.\,(19) three additional partial models, namely OLR, CDDA\textsuperscript{+} and JOLR-DA:

- OLR: In this setting, the partial model only makes use of the NRS\textsubscript{OLR} term, i.e., noise robust sparse orthogonal label regression term as in Eq.\,(6) and ignores the rest of our final model, i.e., data distribution alignment term as defined in Eq.\,(3) as well discriminative MMD constraint terms as defined in Eq.\,(4, 5). This partial model amounts to make use of a particular classifier, i.e., noise robust sparse orthogonal label regression, and enables to quantify the importance of data distribution alignment in DA when contrasted with the baseline JDA;

- CDDA\textsuperscript{+}: In this setting, the regularization term of NRS\textsubscript{OLR} (Eq.\,(6)) is simply replaced by the Nearest Neighbor (NN) predictor. This correspond to our final DA model as defined in Eq.\,(8) and Eq.\,(18) which only make use of the Discriminative MMD constraint terms but without the NRS\textsubscript{OLR} term as defined by Eq.\,(6). In comparison with CDDA, i.e., Close yet Discriminative DA, the partial model already studied in our former DGA-DA method [39] which demonstrated the effectiveness of discriminative force (RF) term in DA. CDDA\textsuperscript{+} includes the improved RF term which adds the RF within the source domain as defined by Eq.\,(5) to the across domain RF defined in Eq.\,(4) as in CDDA. This partial model makes it possible to emphasize the interest of the joint optimization of the improved RF and NRS\textsubscript{OLR} terms in contrasting DOLL-DA with CDDA\textsuperscript{+};

- JOLR-DA: In this setting, the partial model of the proposed method only cares about data distribution alignment as in JDA as well as the newly introduced NRS\textsubscript{OLR} term but ignores the discriminative force as defined in Sect.\,III-B2 and Sect.\,III-B3. In studying JOLR-DA, we aim to highlight: 1) the contribution of NRS\textsubscript{OLR} in regularizing the MMD constraints in contrasting DOLL-DA with JOLR-DA; 2) the effectiveness of NRS\textsubscript{OLR} and the proposed joint optimization strategy, in confronting JOLR-DA with JDA.

- DOLL-DA: This setting correspond to our full final model as defined in Eq.\,(19). It thus contains both CDDA\textsuperscript{+} as defined in Eq.\,(3, 4, 5) and the NRS\textsubscript{OLR} term as defined by Eq.\,(6) in sect.\,\,III-B4.

D. Experimental Results and Discussion

1) Experiments on the CMU PIE Data Set: The CMU PIE dataset is a large face dataset featuring both illumination and pose variations. Fig.\,6 synthesizes the experimental results for DA using this dataset, where top results are highlighted in red color. As expected, without data distribution alignment OLR, with 55.88\% average accuracy, performs worse than the base line JDA with 60.24\% average accuracy. In accounting for the discriminative force in data distribution alignment, CDDA\textsuperscript{+} improves over JDA by 3 points and achieves 63.22\% average accuracy. In adding noise robust sparse orthogonal label regression to JDA, JOLR-DA achieves 69.96\% average accuracy and improves over JDA by 9 points, thereby demonstrating the effectiveness of the NRS\textsubscript{OLR} term. Now our final model, DOLL-DA, with 82.50\% average accuracy, achieves the state of the art performance on this dataset and improves over the baseline JDA by 22 points and the former state of the art DA method, i.e., DGA-DA, by a large margin of 17 points, thereby demonstrates with force the interest of joint optimization of discriminative data distribution terms and the NRS\textsubscript{OLR} term.

2) Experiments on the COIL 20 Dataset: The COIL dataset (see fig.\,5) features the challenge of pose variations between the source and target domain. Fig.\,7 reports the experimental results on the COIL dataset and displays similar patterns as those on the PIE dataset. OLR performs worse than JDA which is improved by CDDA\textsuperscript{+} and JOLR-DA. Finally, DOLL-DA achieves 96.84\% and further improves CDDA\textsuperscript{+} and JOLR-DA by 4 and 3 points, respectively. It is interesting to note that DGA-DA achieves 100\% average accuracy on this dataset and thereby outperforms DOLL-DA by 3.16 points. However, DGA-DA’s outstanding performance is mainly related to the two particularities of the COIL 20 dataset. Indeed, the COIL dataset synthesizes 20 objects as foreground in varying their pose while the background is pure black, therefore each sub-domain contains a single object and is naturally distributed within a specific manifold. Furthermore, the COIL 20 dataset merely contains 20 classes in contrast with 68 classes in the PIE dataset, thereby its classes are much more separated than those in PIE. As a matter of fact, the most simple baseline NN, i.e., Nearest Neighbor, already achieves a high average accuracy of 83.20\% on COIL 20 while it only displays 34.76\% average accuracy on PIE. As a result, in explicitly modeling the hidden data manifold structure through a Laplace graph, DGA-DA performs better label inference than DOLL-DA.

3) Experiments on the Office-Home Dataset: As introduced in DAH[60], Office-Home is a novel challenging benchmark for the DA task. It contains 4 very different domains with 65 object categories, thereby generating 12 different DA tasks. Fig.\,8 synthesizes the performance of the proposed DA methods with DeCAF6 features in comparison with the state of the art methods. Both DAH and DAN are deep DA methods and make use of multi-kernel MMD in aligning the source and target domain distributions. With 43.46\% and 45.54\% average accuracy, respectively, they surpass JDA with a margin up to 8 points. In extending JDA with repulsive force terms, CDDA\textsuperscript{+} slightly improves JDA by 0.19 point. However, thanks to the NRS\textsubscript{OLR} term, JOLR-DA displays 44.48\% average accuracy and improves JDA by 7 points whereas the proposed full model, DOLL-DA, achieves the novel state of the art performance with 48.23\% average accuracy and improves CDDA\textsuperscript{+} by 11.07 points, JOLR-DA by roughly 4 points, and the former state of the art performance achieved by DAH with a margin of 2.69 points.

4) Experiments on the USPS+MNIST Data Set: The USPS+MNIST dataset displays different writing styles between source and target. In Fig.\,9, the left columns of the red vertical bar report the experimental results using shallow...
features, whereas the right columns of the red bar display the results of the methods using deep features. Once more, the proposed DOLL-DA along with its partial models display the same patterns on this large scale benchmark as in the previous experiments. In ignoring the domain shifts, OLR achieves a poor performance. Both CDDA+ and JOLR-DA improve JDA with a large margin. In accounting jointly for the discriminative force and NRS_OLR, DOLL-DA boosts the baseline JDA by 22 points and surpasses DGA-DA, the former state of the art DA method on this dataset, by 4.5 points.

6) Experiments on the Office+Caltech-256 Data Sets: Fig.11 and Fig.12 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., SURF features) are used, respectively.

As can be seen in Fig.12, using shallow features, CDDA+ and JOLR-DA improve the baseline JDA by 2 and 1 points, respectively, thanks to the discriminative force term and NRS_OLR term introduced into the baseline model. Taking them together, our final model DOLL-DA further improves JDA by 4 points and achieves 51.18 average accuracy which is in par with the previous state the art method, namely JGSA, with its 50.58% average accuracy. It is worth noting that, JGSA also suggests aligning data both statistically, discriminatively and geometrically, and corroborates data geometry aware DA approach, and achieves very good performance on this dataset.

Fig.11 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.11, the use of deep features has enabled impressive accuracy improvement over shallow features. Simple baseline methods, e.g., NN, PCA, see their accuracy

Fig. 6: Accuracy% on the PIE Images Dataset.

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

Fig. 8: Accuracy% on the Office-Home Images Dataset.
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 soar from 48.26 to 89.59 for CDDA+, from 47.57 to 88.20 for JOLL-DA, and from 50.78 to 91.65 for DOLL-DA. As for shallow features, CDDA+ and JOLL-DA improve JDA by roughly 3 and 2 points, respectively, while the final model, DOLL-DA, displays the best average accuracy of 92.94% in par with 92.80% displayed by MEDA.

Fig. 9: Accuracy of the USPS+MNIST Images Dataset.

Fig. 10: Accuracy on the SVHN-MNIST Images Dataset.

JOLL-DA. Despite the proposed DOLL-DA displays state of the art performance over 49 DA tasks through 8 datasets except for the COIL dataset where it achieves the second best performance, an important question is how fast the proposed for the repulsive force parameters (sect.IV-E) as well as its sensitivity w.r.t. its hyper-parameters (sect.IV-E1). Additionally, we are curious about how well DOLL-DA performs in changing the base classifier to optimize the classification accuracy of the proposed repulsive force parameters, as well as the DOLL-DA leveraging random initialization (sect.IV-E3) instead of using the base classifier for optimization. Furthermore, in analyzing the generalization capability (sect.IV-E5) of DOLL-DA, we use an unseen target dataset for detail exploration.

5) Sensitivity of the proposed DOLL-DA w.r.t. hyper-parameters: Three hyper-parameters, namely $k$, $\beta$, and $\alpha$, are introduced in the proposed methods.

Dimensionality analysis: $k$ is the dimension of the searched shared latent feature subspace between the source and target domain. In Fig.13, we plot the classification accuracies of the proposed DA method w.r.t different values of $k$ and target domain. As shown in Fig.13, the subspace dimensionality $k$ varies with $k \in \{20, 40, 60, 80, 100, 150, 200, 300, 400\}$, yet the proposed 3 DA variants, namely JOLL-DA and DOLL-DA, remain stable w.r.t. a wide range of $k \in \{40 \leq k \leq 400\}$. It can be seen that both JOLL-DA and DOLL-DA display better robustness than CDDA+ w.r.t. the variation of $k$, thereby suggesting the effectiveness of the NRS-OLR term in the loss function of the proposed DOLL-DA model. Obviously, the larger $k$ the better the shared subspace can afford complex data distributions, but the cost of increased computation complexity is highlighted in sect.III-E on time complexity analysis. In our experiments, we set $k = 300$ to balance the efficiency and accuracy.

Sensitivity of $\alpha$ and $\beta$: $\alpha$ and $\beta$ as defined in Eq.(19) are the major hyper-parameters of the proposed DOLL-DA. While $\alpha$ aims to regularize the projection matrix $A$ to avoid over-fitting the chosen shared feature subspace w.r.t. both source and target domain data, $\beta$ as expressed in Eq.(6) controls the dimensionality of class dependent data manifold in the searched shared feature subspace, or in other words the sparsity level of the linear combination of the projected
features to regress the class label. We study the sensitivity of the proposed DOLL-DA method with a wide range of parameter values, i.e., $\alpha = (0.001, 0.01, 0.1, 1, 10, 20, 50)$ and $\beta = (0.05, 0.1, 1, 5, 10, 100, 200)$. We plot in Fig.14 the results on $D \rightarrow W$, $C \rightarrow D$ and PIE-27 $\rightarrow$ PIE-5 datasets on the proposed DOLL-DA with $k$ held fixed at 300. As can be seen from Fig.14, the proposed DOLL-DA displays its stability as the resultant classification accuracies remain roughly the same despite a wide range of $\alpha$ and $\beta$ values.

Fig. 13: 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 JOLR-DA; (c) accuracy w.r.t. subspace dimension $k$ of DOLL-DA. Three datasets are used, i.e., COIL1, COIL2 and PIE.

2) Convergence analysis: In Fig.15, we further perform convergence analysis of the proposed DOLL-DA along with its partial models, i.e., CDDA+ and JOLR, using the DeCAF6 features on the Office+Caltech datasets and pixel value features on the PIE dataset. We aim to disclose how fast the proposed methods achieve their best performance w.r.t. the number of iterations $T$. Fig.15 reports 6 cross DA experiments ($C \rightarrow A$, $D \rightarrow W$ ... PIE-27 $\rightarrow$ PIE-5) with the number of iterations $T = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$. As shown in Fig.15, CDDA+, JOLR-DA and DOLL-DA converge within 3~5 iterations during optimization, but JOLR-DA and DOLL-DA seem to converge even faster with a better accuracy, thanks to the NRS_OLR term introduced in our DA model.

Fig. 15: Convergence analysis using 6 cross-domain image classification tasks on Office+Caltech256 and PIE datasets. (accuracy w.r.t. #iterations)

3) Impact of the base classifier: The repulsive force term across domain as formulated in sect.III-B2 requires using pseudo labels for the target domain data. Therefore, the quality of these pseudo labels could have much impact on the effectiveness of data discriminativeness. In sect.III-C, our model is initialized using pseudo labels in solving Eq.(8), which boils down to JDA. The inference of the pseudo labels on the target domain data requires a base classifier trained using the labeled source domain data. We test the sensitivity of the proposed DOLL-DA w.r.t. the base classifier using two popular classifiers, i.e., NN and SVM. Fig.16 shows that DOLL-DA-NN(92.94%) and DOLL-DA-SVM(91.51%) achieve almost the same performance. This result suggests that our repulsive force term in the proposed DOLL-DA displays a certain level of robustness w.r.t. the choice of the base classifier.

4) Random Label Initialization: Going one step further w.r.t. to the experiment in sect.IV-E3, an interesting question is how DOLL-DA behaves with randomly initialized pseudo labels at its first iteration as well as its convergence efficiency in this specific experiment setting.

In this setting, DOLL-DA makes use of randomly initialized labels for the target domain data instead of solving Eq.(8) at its first iteration. As shown in Fig.16, DOLL-DA still achieves 90.01% accuracy on the Office-Caltech256 dataset, thus only slightly below DOLL-DA-NN, and converges on average at 3.08 average iterations. This result further supports the robustness of the designed repulsive force term regularized by the NRS_OLR term in the search of the optimized shared features subspace.

| $C \rightarrow A$ | $C \rightarrow W$ | $C \rightarrow D$ | $A \rightarrow C$ | $A \rightarrow W$ | $A \rightarrow D$ | $W \rightarrow C$ | $W \rightarrow A$ | $W \rightarrow D$ | $D \rightarrow C$ | $D \rightarrow A$ | $D \rightarrow W$ | Avg. acc. |
|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|-------------|
| DOLL-DA-NN       | 93.66            | 93.98            | 91.71            | 89.87            | 91.86            | 91.72            | 86.46            | 93.11            | 100.00          | 86.78            | 93.59          | 91.00          | 92.94       |
| DOLL-DA-SVM      | 93.91            | 93.32            | 92.76            | 85.50            | 91.86            | 91.27            | 86.46            | 93.11            | 99.41            | 86.22            | 91.54          | 91.51          | 91.51       |
| Rand Init        | 93.22            | 93.47            | 95.98            | 96.59            | 91.38            | 90.21            | 86.46            | 93.11            | 99.41            | 96.22            | 91.54          | 91.51          | 91.51       |
| #Iter (ts)       | 2.00             | 2.00             | 2.00             | 2.00             | 2.00             | 2.00             | 2.00             | 3.00             | 2.00             | 4.00             | 3.00             | 3.00          | 3.00        |

Fig. 16: Convergence analysis using 12 cross-domain image classification tasks on Office+Caltech256 datasets with DeCAF6 Features. (accuracy w.r.t. #iterations)

5) Stability w.r.t. target domain data: We benchmark the stability of the proposed DOLL-DA w.r.t. the quantity of target domain data used for training. Specifically, we carry out two additional experiments for 2 DA tasks using the Office-Caltech256 dataset, i.e., $W \rightarrow A$, $D \rightarrow W$ keeping 5% (10%, 30%, 50%, 70%, resp.) of target domain data from being used as auxiliary data in training. Results are reported in Fig.17. As can be seen there, when all target domain data are used in training the proposed DOLL-DA, i.e., column 0.00%, DOLL-DA displays 93.11% and 100% accuracy for $W \rightarrow A$ and $D \rightarrow W$ DA tasks, respectively. Now when more and more target domain data are kept from being used in training, passing from 5% through 70%, DOLL-DA proves quite stable on the $D \rightarrow W$ task but decreases constantly to reach 79.96% on the $W \rightarrow A$ task. However, this result still proves the usefulness of the proposed DA method when only 30% target domain data are used as auxiliary data in training, given the fact that the baseline NN only displays 73.07% accuracy as shown in Fig.11. The performance difference between these
two DA tasks can be explained by their inherent difficulties. The DA task $D \rightarrow W$ is to generalize a classifier trained from the source domain, i.e., labeled images in the DSLR domain, thus with much background, to the target domain, i.e., images in the Webcam domain, much simpler because devoid of the background, and the simple baseline NN already achieves 97.97% accuracy as shown in Fig.11. On the other side, the DA task $W \rightarrow A$ does exactly the contrary and needs to generalize a classifier trained from the source domain, i.e., labeled images in the Webcam domain, thus without background, to a much more complicated target domain, i.e., images in the Amazone domain with arbitrary background, and the simple baseline NN only achieves 73.07% accuracy as shown in Fig.11.

| $D \rightarrow W$ | Accuracy |
|------------------|----------|
| Unseen 0.00%     | 93.11    |
| 5.00%            | 92.59    |
| 10.00%           | 86.07    |
| 30.00%           | 85.85    |
| 50.00%           | 84.32    |
| 70.00%           | 79.96    |

| $W \rightarrow A$ | Accuracy |
|-------------------|----------|
| Unseen 0.00%      | 100      |
| 5.00%             | 98.75    |
| 10.00%            | 98.75    |
| 30.00%            | 98.75    |
| 50.00%            | 98.75    |
| 70.00%            | 98.75    |

Fig. 17: Unseen data sensitivity using 2 cross-domain image classification tasks on Office+Caltech256 datasets with DeCAF6 Features.

V. CONCLUSION

We have proposed in this paper a novel unsupervised DA method, namely Discriminative Noise Robust Sparse Orthogonal Label Regression-based Domain Adaptation (DOLL-DA), which simultaneously optimizes the three terms of the upper error bound of a learned classifier on the target domain in aligning discriminatively data distributions through a repulse force term while orthogonally regressing data labels within the shared feature subspace. Furthermore, the proposed model explicitly accounts for data outliers to avoid negative transfer and introduces the property of sparsity when regressing data labels. Comprehensive experiments using the standard benchmark in DA show the effectiveness of the proposed method which consistently outperform state of the art DA methods. Future work includes embedding of the proposed DOLL-DA into the paradigm of deep learning and considers the setting of online learning where target domain data only arrives sequentially one after another.

REFERENCES

[1] 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.

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In Advances in neural information processing systems, pages 137–144, 2007.

[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.

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In Advances in Neural Information Processing Systems, pages 343–351, 2016.

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

[6] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In Advances in Neural Information Processing Systems, pages 3733–3742, 2017.

[7] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence, 39(9):1853–1865, 2017.

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

[9] Zhengming Ding and Yun Fu. Robust multiview data analysis through collective low-rank subspace. IEEE Trans. Neural Netw. Learning Syst., 29(5):1986–1997, 2018.

[10] 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.

[11] 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.

[12] Basura Fernando, Amaury Habrard, Marc Sebban, and Timme 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.

[13] Simone Fiori. Formulation and integration of learning differential equations on the stiefel manifold. IEEE transactions on neural networks, 16(6):1697–1701, 2005.

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 17(1):2096–2030, 2016.

[15] Muhammad Ghifary, David Baldiuzzi, 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.

[16] 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.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[18] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. Advance in neural information processing systems, pages 513–520, 2007.

[19] Mehrtash Harandi, Mathieu Salzmann, and Richard Hartley. Dimensionality reduction on spd manifolds: The emergence of geometry-aware methods. IEEE transactions on pattern analysis and machine intelligence, 2017.

[20] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Learning an invariant hilbert space for domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3845–3854, 2017.

[21] Samitha Herath, Mehrtash Tafazzoli Harandi, and Fatih Porikli. Learning an invariant hilbert space for domain adaptation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 3956–3965, 2017.

[22] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 1989–1998, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

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

[24] 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.

[25] Alireza Karbalayeh, Xiaoning Qian, and Edward R Dougherty. Optimal bayesian transfer learning. IEEE Transactions on Signal Processing, 66(14):3724–3739, 2018.
[26] 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.

[27] 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. 4, 9, 10

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998. 9, 10, 12

[29] Jingjing Li, Mengmeng Jing, Ke Lu, Lei Zhu, and Heng Tao Shen. Locality preserving joint transfer for domain adaptation. IEEE Transactions on Image Processing, 28(12):6103–6115, 2019. 10

[30] Jian Liang, Ran He, Zhenan Sun, and Tieniu Tan. Aggregating randomized clustering-promoting invariant projections for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. 1, 9

[31] Youfa Liu, Weiping Tu, Bo Du, Lefei Zhang, and Dacheng Tao. Homomorphic component analysis for domain adaptation. IEEE Transactions on Image Processing, 29:1074–1089, 2019. 3

[32] Mingsheng Long, Yue Cao, Jiannin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In ICML, pages 97–105, 2015. 4, 10

[33] Mingsheng Long, Jiannin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Y. Transfer feature learning with joint distribution adaptation. In Proceedings of the IEEE International Conference on Computer Vision, pages 2200–2207, 2013. 1, 3, 9, 10

[34] Mingsheng Long, Jiannin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Y. 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. 10

[35] Mingsheng Long, Han Zhu, Jiannin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, pages 2208–2217, 2017. 1, 4

[36] Hao Lu, Chunhua Shen, Zhiqiu Cao, Yang Xiao, and Anton van den Hengel. An embarrassingly simple approach to visual domain adaptation. IEEE Transactions on Image Processing, 2018. 1, 3

[37] Ying Lu, Lingkun Luo, Di Huang, Yunhong Wang, and Liming Chen. Knowledge transfer in vision recognition: A survey. ACM Comput. Surv., 53(2), April 2020. 1

[38] Lingkun Luo, Liming Chen, Shiqiang Hu, Ying Lu, and Xiaofang Wang. Discriminative and geometry-aware unsupervised domain specialization. CoRR, abs/1712.10042, 2017. 1

[39] Lingkun Luo, Liming Chen, Shiqiang Hu, Ying Lu, and Xiaofang Wang. Discriminative and geometry-aware unsupervised domain adaptation. IEEE Transactions on Cybernetics, 2020. 1, 2, 3, 4, 6, 10, 11, 12

[40] 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, 9, 10

[41] Feiping Nie, Jianjun Yuan, and Heng Huang. Optimal mean robust weights for deep domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10):1345–1359, 2010. 1

[42] Siinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Discriminative and geometry aware unsupervised domain adaptation. In International conference on machine learning, pages 1062–1070. 2014. 8

[43] Siinno 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. 3, 10

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

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

[46] 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

[47] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jiannin Wang. Multi-adversarial domain adaptation. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018. 4

[48] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. 9, 10, 12

[49] 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. 10

[50] 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

[51] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In Advances in Neural Information Processing Systems, pages 2110–2118, 2016. 10

[52] 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, 10

[53] 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. 3, 10

[54] Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. IEEE Transactions on Knowledge and Data Engineering, 22(7):929–942, 2010. 1

[55] Baochen Sun and Kate Saenko. Return of frustratingly easy domain adaptation. In AAAI, volume 6, page 8, 2016. 1, 3, 10

[56] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In European Conference on Computer Vision, pages 443–450. Springer, 2016. 4

[57] 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

[58] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In Computer Vision and Pattern Recognition (CVPR), volume 1, page 4, 2017. 1, 4, 10

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

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

[61] Hemantam Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethu Raman Pancharathan. Deep hashing network for unsupervised domain adaptation. arXiv preprint arXiv:1706.07522, 2017. 9, 10, 11

[62] 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, Quebec City, Quebec, Canada., pages 2099–2105, 2014. 10

[63] Jindong Wang, Wenhui Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In 2018 ACM Multimedia Conference on Multimedia Conference, pages 402–410. ACM, 2018. 4

[64] Yong Xu, Xiaohao 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, 10

[65] 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, 10

[66] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for transfer subspace learning. In Advances in Neural Information Processing Systems, pages 7285–7298, 2019. 4
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, Ecullly, France, as a Professor in 1998, where he leads an advanced research team on Computer Vision, Machine Learning and Multimedia. From 2001 to 2003, he also served as Chief Scientific Officer in a Paris-based company, Aviías, 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, Ecole Centrale de Lyon from 2007 through 2016. His current research interests include computer vision, machine learning, image and multimedia with a particular focus on robot vision and learning since 2016. Liming has over 300 publications and successfully supervised over 40 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 5 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. He has over 150 publications and successfully supervised over 15 PhD students. Now, he is a full professor in Shanghai Jiao Tong University. His research interests include data fusion technology, image understanding, and nonlinear filter.