Iran Journal of Computer Science (2023) 6:345–364
https://doi.org/10.1007/s42044-023-00144-x

RESEARCH

Unsupervised domain adaptation via transferred local Fisher discriminant analysis

Mozhdeh Zandifar[1] · Samaneh Rezaei[2] · Jafar Tahmoresnezhad[2]

Received: 17 December 2022 / Accepted: 14 April 2023 / Published online: 30 April 2023

Abstract
Domain adaptation in machine learning and image processing aims to benefit from gained knowledge of the multiple labeled training sets (i.e. source domain) to classify the unseen test set (i.e. target domain). Therefore, the major issue emerges from dataset bias where the source and target domains have different distributions. In this paper, we introduce a novel unsupervised domain adaptation method for cross-domain visual classification. We suggest a unified framework that reduces both statistical and geometrical shifts across domains, referred to as unsupervised domain adaptation via transferred local Fisher discriminant analysis (TLFDA). Specifically, TLFDA projects data into a shared subspace to minimize the distribution shift between domains and simultaneously preserves the discrimination across different classes. TLFDA maximizes the between-class separability and preserves the within-class local structure in form of an objective function metric. The objective function is solved effectively in closed form. Broad experiments demonstrate that TLFDA significantly outperforms many state-of-the-art domain adaptation methods on different cross-domain visual classification tasks.

Keywords Transfer learning · Unsupervised domain adaptation · Dimensionality reduction · Fisher discriminant analysis · Locality preserving projection · Bregman divergence

1 Introduction

Recently, in the machine learning and pattern recognition fields, enormous amounts of data, e.g., images, videos, and texts are emerging where the traditional supervised machine learning methods need to label data for each gallery or corpus [1, 2]. In fact, in most existing applications, there are not sufficient labeled data to classify the new domains while the manual labeling of the unlabeled instances is immensely intricate and expensive. Thus, the vital significance of using other existing related labeled domains to classify the new visual domains has drawn more consideration during the last few years. However, the classification results are often poor when the trained classifiers on the available labeled samples directly are used to classify the new unlabeled instances with various distributions. For example, imagine that we are to develop an iPhone app to identify the car-captured images via a phone's camera while there are no labeled images. In this case, the trained model would not work appropriately since the training and test images have various expressions, postures, and lighting conditions which means different distributions [3].

The challenge of exploiting other related domains with different distributions or feature spaces to classify new tasks presents the domain shift problem. To address the problem, a variety of solutions have been developed named transfer learning (TL) and domain adaptation (DA) which are the improvement of the learning paradigms in a new task through the transferring of knowledge from related tasks that have been already learned.

Generally, DA and TL techniques are categorized into two different settings. The first setting is called semi-supervised domain adaptation in which a few parts of the target domain are labeled and the rest is unlabeled. In the second setting, called unsupervised domain adaptation, there is no labeled sample in the target domain [4]. However, in both settings, the source and target data often have different marginal and conditional distributions. Some basic DA methods, only consider the marginal distribution disparity between domains

✉ Jafar Tahmoresnezhad
j.tahmores@it.uut.ac.ir

[1] Faculty of IT and Computer Engineering, Urmia University of Technology, Urmia, Iran

[2] Department of IT and Computer Engineering, Urmia University of Technology, Urmia, Iran

and often ignore the conditional distribution discrepancy. Unlike previous methods, our proposed approach exploits the geometry of the data manifold to minimize both the marginal and the conditional distribution differences.

Learning invariant features while the distribution of the source and target domains are different is decisive. However, most of the traditional dimensionality reduction approaches like Fisher linear discriminant analysis (FDA) [5] and locality preserving projections (LPP) [6], perform poorly to encounter domain shift problems either in original or low-dimensional spaces. Thus, we are to develop a new dimensionality reduction algorithm to address the domain shift issue.

on the other hand, Bregman divergence (BD) [7] is a nonlinear measure to minimize the distance between distributions using the kernel density estimation (KDE) technique [8]. Specifically, we extend the nonlinear Bregman divergence to measure the discrepancy of marginal and conditional distributions and integrate it with local Fisher discriminant analysis (LFDA) [9] to create a new feature representation that is efficacious and robust against considerable distribution divergence. LFDA effectively combines the ideas of FDA and LPP to maximize the between-class separability and preserve the within-class local structure, simultaneously. Furthermore, the Bregman divergence can transfer the gained knowledge from training to test sets by minimizing the mismatch across the marginal and conditional distributions of them.

In this work, to tackle with unsupervised DA problem, we propose a novel domain adaptation approach called unsupervised domain adaptation via transferred local Fisher discriminant analysis (TLFDA), which projects the source and target data into a common subspace such that both the marginal and conditional distribution discrepancies of domains are minimized. Moreover, TLFDA exploits the Bregman divergence to measure the distribution difference, which enables transferring the knowledge from training samples to test ones. Furthermore, TLFDA maximizes the between-class separability of source and target domains and preserves the within-class local structure in a low-dimensional subspace. TLFDA considers both the discriminative information of marginal instances in various classes and the local geometry of instances in each class.

Contributions: The contributions of our TLFDA are listed as follows.

1. TLFDA mitigates the joint marginal and conditional distribution discrepancies across the source and target domains via Bregman divergence.
2. TLFDA introduces a novel dimensionality reduction method for the domain shift problem where its idea originates from joint FDA and LPP.
3. TLFDA predicts the pseudo-labels of the target data based on a conscious estimate via a trained model on source data.
4. TLFDA benefits from an iterative approach to refining the pseudo-labels of target samples.

Organization of the paper: The next section provides a review of related work. We then represent our proposed method in Sect. 3. In Sects. 4 and 5, we provide the experimental results and discussion. Finally, the paper is concluded in Sect. 6, and future works are included.

2 Related work

The existing DA methods to tackle the problem of domain shift are organized into three adaptation categories [10, 11]: (1) instance-based methods, (2) model-based methods, and (3) feature-based methods.

The instance-based approaches [12, 13] are to assign less importance to the irrelevant source instances to reduce the distribution discrepancy across the source and target domains. Landmark selection [14] is one of the instance-based methods, which benefits from max mean discrepancy (MMD) [15] to select a subset of source samples that obey the same distribution as target samples. In the other words, the major focus of the landmark selection method is to conjoin the source data with the target one using the discovered landmarks. LSSA [16] is another instance-based method selecting a subset of instances as landmarks and projects the source and target data into a latent subspace in a non-linear procedure considering the selected landmarks. LSSA uses the subspace alignment to adapt the unaligned domains by learning a non-linear mapping function.

The model-based methods [17, 18] center around the notion of adaptive classifier design to have a robust model to cope with the distribution mismatch across domains by transferring the model parameters from a source-made model to another target model. Domain adaptation machine (DAM) [19] benefits from a set of auxiliary/source classifiers that are trained with the labeled samples from many source domains to learn a robust target classifier. Adaptation regularization-based transfer learning (ARTL) [20] is a novel method that reduces structural risk and jointly minimizes the marginal and conditional distribution difference between domains. Also, ARTL maximizes the manifold consistency to tackle unsupervised domain adaptation.

The feature-based methods [21–30] change the feature representation of the source and target domains to bring closer the marginal and conditional distributions of domains, jointly. Joint distribution adaptation (JDA) [21] creates a novel feature representation using a principled dimensionality reduction technique that is robust against distribution

shift. Low-rank and sparse representation (LRSR) [22] preserves the inherent geometric data structure via low-rank constraint and/or sparse representation in an embedded subspace. LRSR geometrically aligns the source and target data through both the low-rank and sparse constraints such that the source and target data are interleaved within a new shared feature subspace. Visual domain adaptation (VDA) [23] uses the joint domain adaptation and transfer learning to deal with the problem of domain shift. VDA discriminates various classes in an embedded representation via condensed domain-invariant clusters. Close yet discriminative domain adaptation (CDDA) [31] is a novel framework that constructs a common feature representation with the following two properties. First, the difference across the source and target domains is measured in terms of joint marginal and conditional probability distributions via maximum mean discrepancy. Second, CDDA discriminates data using the inter-class repulsive force. Coupled local–global adaptation (CLGA) [24] globally adapts the marginal and conditional distribution disparities. CLGA builds a graph to minimize the distances across the sample pairs in the same class manifold through the different domain manifolds and to maximize the distances across the sample pairs in the same domain manifold with different class manifolds. Domain invariant and class discriminative representations (DICD) [25] jointly matches the marginal and conditional distributions in a latent subspace. DICD discriminates classes by increasing the intra-class density and also decreasing the inter-class dispersal. Discriminative and geometry aware domain adaptation (DGA-DA) [26] defines a repulsive form term to discriminate the latent feature space. DGA-DA infers the labels using the geometric structures of explored data through label smoothness and geometric structure consistencies. Transductive transfer learning for image classification (TTLC) [27], as a state-of-the-art domain adaptation method, globally adapts the marginal and conditional distributions in two respective low-dimensional subspaces. TTLC regulates the distances across sample pairs in both domains to discriminate various classes. At last, TTLC locally aligns both latent subspaces. Domain adaptation with geometrical preservation and distribution alignment (GPDA) [28] preserves both the statistical and the geometrical properties of domains in a unified framework. Firstly, GPDA tries to preserve the statistical properties of data via the nonnegative matrix factorization model [32]. Then, GPDA preserves the geometrical structure of data through graph dual regularization in the nonnegative matrix factorization framework. Also, the marginal and conditional distribution disparities are aligned in the mentioned framework. Unified cross-domain classification method via geometric and statistical adaptations (UCGS) [29] minimizes the structural risk on source data, at first. Also, UCGS uses MMD for statistical adaptation and the Nystrom method for geometrical adaptation in a unified framework. Feature selection-based visual domain adaptation (FSVDA) [30] uses the particle swarm optimization (PSO) [33] algorithm to select the most relevant feature subsets across both domains. To evaluate the effectiveness of each subset, FSVDA uses the manifold embedded distribution alignment (MEDA)'s function [34] as its fitness function.

Recently, a great effort has been dedicated to deep DA methods. Deep methods for extracting features via hidden layers need a high amount of training data. Venkateswara et al. proposed a domain adaptive hashing network (DAH) [35] to assign unique hash codes for source and target domains. Manifold Aligned Label Transfer for Domain Adaptation (MALT-DA) [36] uses a densely connected architecture (DenseNet) [37] to learn better deep features on the source domain. MALT-DA aligns features across domains through two methods including Adaptive Batch Normalization (ABN) [38], and subspace alignment via LPP. Following, MALT-DA clusters the features into variant groups. MALT-DA compares the labels which are made via the cluster-matching process and the labels which are hypothesized via the network. Then, the samples with matching labels are used as training data for the adaptation method. Deep methods have more time complexity for the training phase whereas TLFDA with a convex time complexity can be preferred to deep methods.

As a result, our current research in TL has focused on the third category, the feature-based adaptation approach, which looks for a common feature representation across the source and target domains. In summary, our main contribution is a new dimensionality reduction-based method that combines the ideas of FDA and LPP to minimize the distance between the marginal and conditional distributions of source and target data to enable effective transfer learning. TLFDA unlike most of the previous works decreases the marginal and conditional discrepancies via Bregman divergence. Moreover, TTLC maps source and target samples into respective subspaces but TLFDA maps samples into a common subspace. We apply our new method to four real-world applications in a transfer learning setting to demonstrate its outstanding performance.

3 Proposed method

In this section, we describe a precise description of TLFDA to deal with the unsupervised domain shift problem. In the case of domain shift problems, considering the distribution nonconformity between the source and target domains is vital for achieving the desired results for methods that are based on FDA and LPP criteria. To this end, in this work, we represent a novel dimensionality reduction framework for the domain shift problem where its idea originates from joint FDA, LPP, and Bregman divergence. Our contribution in this work is to

find a common subspace where the marginal and conditional distribution discrepancies of domains are jointly minimized and the local structure of data is well preserved.

To this end, we suppose $D = < X, P(x) >$ as a domain that consists of the instances in feature space X with marginal probability distribution $P(x)$. Moreover, we consider $T = < Y, f(x) >$ as a task for domain D that consists of a label set Y and a prediction function $f(x)$. Note that $f(x)$ can be interpreted as a conditional probability distribution, i.e., $P(Y|x)$. In this paper, an unsupervised domain shift problem with $D_S = \{(x_1, y_1), \ldots, (x_{n_s}, y_{n_s})\}$ as the labeled source domain and $D_T = \{x_{n_{s+1}}, \ldots, x_{n_s+n_t}\}$ as an unlabeled target domain is addressed where n_s and n_t are the number of the source and target samples, respectively.

3.1 Dimensionality and divergence reduction

The major idea behind our proposed approach is to discover an optimal couple of projections and classification for the source and target instances in a way that the distribution divergence between domains is decreased. A feasible way to attain this objective is to use dimensionality decrement methods (e.g., LPP and FDA). However, contrary to the success of such methods, they cannot guarantee the model's efficiency against the problem of domain shift. Thus, as well as exploiting the dimensionality reduction approaches, we are to consider domain adaptation settings to build a robust model against distribution shift. In this section, at first, we formulate the classical FDA and LPP, and afterward, we introduce the classical LFDA [9], which is a combination of dimensionality reduction methods.

The main objective of FDA is to express one dependent variable as a linear combination of other variables. To this end, FDA extracts the new features of the domain according to the linear combination of available features. In fact, FDA attempts to maximize the class-separate degree. Therefore, FDA incorporates the following two criteria. (1) FDA maximizes the between-class scatter matrix, and (2) minimizes the within-class scatter matrix, such that the samples in the embedded subspace have maximum discrimination. Let $x_i \in R^D (i = 1, 2, \ldots, n)$ be a D-dimensional sample and $y_i \in 1, 2, \ldots, c$ be the related class label of ith sample, where n is the number of instances and c is the number of classes. Let n_l be the number of instances in class l, where $\sum_{l=1}^{c} n_l = n$. Mathematically, the between-class scatter matrix is given by $S^{(b)} = \sum_{l=1}^{c} n_l(\mu_l - \mu)(\mu_l - \mu)^{\mathrm{T}}$ and the within-class scatter matrix is $S^{(w)} = \sum_{l=1}^{c} \sum_{i:y_i=l} (x_i - \mu_l)(x_i - \mu_l)^{\mathrm{T}}$, where the inner sigma stands for the summation over i such that $y_i = l$, μ_l is the mean of instances in class l, and μ is the mean of all instances.

Thus, FDA subspace is given by $\mathrm{argmax}_J tr(J^{\mathrm{T}} S^{(b)} J)/tr(J^{\mathrm{T}} S^{(w)} J)$ subject to $J^{\mathrm{T}} J = I$. Therefore, FDA transformation matrix J is defined as follows:

$$J_{\mathrm{FDA}} = \mathrm{argmax}_J \left[\mathrm{tr}^{-1} \left(J^{\mathrm{T}} S^{(b)} J \right) \mathrm{tr} \left(J^{\mathrm{T}} S^{(w)} J \right) \right], \tag{1}$$

where $\mathrm{tr}^{-1}(.)$ is the inverse of matrix trace. The intuition behind maximizing J_{FDA} is to learn a projection matrix $J \in R^{D \times d}$ to transform data from the original feature space composed of D features into a low dimensional subspace with d features (i.e., $d < D$).

LPP as another dimensionality reduction technique exploits an undirected graph indicating the neighbor relations of pairwise samples to preserve the local geometry of data. Also, LPP optimally approximates the eigenfunctions of the Laplace Beltrami operator over the data manifold, linearly. The weight between instances $\vec{x}_i$ and $\vec{y}_j$ is calculated via $E_{ij} = \exp(-\|\vec{x}_i - \vec{y}_j\|^2/t)$ for the same class samples, and is considered $E_{ij} = 0$ in other cases. The LPP transformation matrix J_{LPP} is defined as follows:

$$\begin{aligned} J_{\mathrm{LPP}} &= \frac{1}{2} \sum_{i,j=1}^{n} \left(\left(J^{\mathrm{T}} x_i - J^{\mathrm{T}} x_j \right)^{\mathrm{T}} \left(J^{\mathrm{T}} x_i - J^{\mathrm{T}} x_j \right) \right) E_{ij} \\ &= 2tr \left(J^{\mathrm{T}} X \left(W - E \right) X^{\mathrm{T}} J \right), \end{aligned} \tag{2}$$

where W is a diagonal matrix with $W_{ii} = \sum_{j=1}^{n} E_{ji}$ and J_{LPP} is a transformation matrix that transforms samples into a latent subspace.

The performance of FDA degrades dramatically where the instances in a class are from several distinct clusters. In fact, it causes via the globality during the within-class and between-class scatters evaluation. Therefore, whenever the samples in various classes are close in the original high-dimensional space R^D, LPP with its unsupervised nature can also overlap them. To dominate these problems, we introduce a novel idea from a combination of FDA and LPP. Since the various classes might come from different distributions, they are treated differently, and thus, there is the dissimilarity among them. Therefore, we maximize the borders across various classes as much as possible. In fact, we introduce local Fisher discriminant analysis (LFDA) as a novel linear dimensionality reduction approach which effectively combines the ideas of FDA and LPP. In fact, having an analytical form of the embedding transformation, the projection matrix can be easily computed just by solving a generalized eigenvalue problem. The LFDA transformation matrix J_{LFDA} is defined as follows:

$$J_{\mathrm{LFDA}} = \mathrm{argmax}_J \left[tr \left(J^{\mathrm{T}} \tilde{S}^{(b)} J \right)^{-1} J^{\mathrm{T}} \tilde{S}^{(w)} J \right], \tag{3}$$

where $\tilde{S}^{(w)}$ can be defined as

$$\tilde{S}^{(w)} = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{n_{y_i}}\tilde{P}_i^{(w)} \tag{4}$$

and n_{y_i} is the number of instances in which the sample x_i belongs and $\tilde{P}_i^{(w)}$ is the pointwise local within-class scatter matrix around x_i and is defined as follows:

$$\tilde{P}_i^{(w)} = \frac{1}{2}\sum_{j:y_j=y_i}^{n} E_{ij}(x_j - x_i)(x_j - x_i)^{\mathrm{T}}. \tag{5}$$

Accordingly, minimizing $\tilde{S}^{(w)}$ corresponds to minimizing the weighted sum of the pointwise local within-class scatter matrices over all instances. Moreover, $\tilde{S}^{(b)}$ can be defined in a similar way as follows:

$$\tilde{S}^{(b)} = \frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{1}{n_{y_i}}\right)\tilde{P}_i^{(w)} + \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{n_{y_i}}\tilde{P}_i^{(b)}, \tag{6}$$

where $\tilde{P}_i^{(b)}$is the pointwise between-class scatter matrix around x_i and is expressed as follows:

$$\tilde{P}_i^{(b)} = \sum_{j:y_j\neq y_i}(x_j - x_i)(x_j - x_i)^{\mathrm{T}}. \tag{7}$$

Note that $\tilde{P}_i^{(b)}$ does not include the localization factor $E_{i,j}$. Moreover, Eq. 6 mentions that maximizing $\tilde{S}^{(b)}$ corresponds to decreasing the weighted sum of pointwise local within-class scatter matrices and maximizing the sum of pointwise between-class scatter matrices. Therefore, eigenvalue decomposition of $\mathrm{tr}^{(-1)}(\tilde{S}^{(w)}\tilde{S}^{(b)})$ is used for achieving the solution of J_{LFDA} as an optimization problem. However, the eigenvectors corresponding to d largest eigenvalues construct the mapping matrix J. Despite LFDA efficiency, it cannot minimize the distribution diversity across the source and target domains. Hence, we are to find a solution to adapt the distribution diversity across domains.

Therefore, the major issue is to reduce the distribution mismatches across the source and target domains by precisely reducing the empirical distance measure. Measuring the distance across distributions by the parametric criteria requires expensive distribution calculation. Thus, we utilize a nonlinear distance measure, referred to as Bregman divergence. Bregman divergence measures the distribution diversity of drawn samples from different domains in a projected subspace. In fact, the Bregman divergence is able to transfer the gained knowledge from training sets to test ones by reducing the distance across the distributions of training and test samples.

The Bregman distance is a generalization of a wide range of distance functions (e.g., Mahalanobis distance [39], square root, and Kullback–Leibler divergence [40]), and is capable to explore the nonlinear correlations of data features. Many Bregman divergence applications cause recent advances in machine learning.

Definition 3.1 (*Bregman divergence*). Given a strictly convex function f on Ω, the Bregman divergence corresponding to f is defined as:

$$BD(x, y) = f(x) - f(y) - (x - y)\bigtriangledown f(y), \tag{8}$$

where $\bigtriangledown f$ represents the gradient vector of f.

The defined Bregman divergence in (3) can be described as the discrepancy across the value of the convex function at x and its first-order Taylor expansion at y, or equivalently the remainder term of the first-order Taylor expansion of f at y. Indeed, the Bregman divergence reduces to a well-known loss function according to the choice of the convex function f. For example, if $f = x^2$, then we have $BD(x, y) = (x - y)^2$, and clearly its square root is a metric. Thus, the functional Bregman divergence $BD(.,.)$ is expressed as

$$\begin{aligned} BD(P_{\mathrm{S}}, P_T) &= \int (P_{\mathrm{S}}(\vec{y}) - P_T(\vec{y}))^2\, d\vec{y} \\ &= \int \left(P_{\mathrm{S}}(\vec{y})^2 - 2P_{\mathrm{S}}(\vec{y})P_T(\vec{y}) + P_T(\vec{y})^2\right) d\vec{y}, \end{aligned} \tag{9}$$

where P_{S} and P_T are the probability density functions (PDFs) of the source and the target data in latent subspaces, respectively. Therefore, Bregman divergence measures the distribution differences of the source and target domains in the latent subspace.

Using the kernel density estimation technique, the densities are estimated in the latent subspace. Therefore, the density is estimated at each point $y \in R^d$ as the sum of kernels between $\vec{y}$ and other points $\vec{y}_i$ as follows:

$$p(\vec{y}) = \left(\frac{1}{n}\right)\mathbf{G}_{\sum}(\vec{y} - \vec{y}_i), \tag{10}$$

where n is the number of samples and $\mathbf{G}_{\sum}(.)$ is the d-dimensional Gaussian kernel with the covariance matrix $\sum$.

3.2 Distribution adaptation using Bregman divergence

The most important challenge in domain adaptation is to decrease the divergence across the source and target domains. BD preserves the geometric structure of data and just minimizes the distribution divergence across domains. However,

aligning both the marginal and the conditional distributions is very effective for robust domain adaptation. To measure the discrepancy across the marginal distributions of the source and target domains, we employ BD as follows:

$$\begin{aligned}
&\mathrm{Dist}^{\mathrm{marginal}}(D_{\mathrm{S}}, D_T) \\
&= \left[\int \left(\frac{1}{n_{\mathrm{s}}} \sum_{i=1}^{n} \mathbf{G}_{\sum_1} (\vec{y} - \vec{y}_i) \right)^2 d\vec{y} \right] \\
&+ \left[\int \left(\frac{1}{n_{\mathrm{t}}} \sum_{j=n_{\mathrm{s}}+1}^{n} \mathbf{G}_{\sum_2} (\vec{y} - \vec{y}_i) \right)^2 d\vec{y} \right] \\
&- \left[\left(\frac{1}{n_{\mathrm{s}} n_{\mathrm{t}}} \sum_{i=1}^{n} \sum_{j=n_{\mathrm{s}}+1}^{n} \mathbf{G}_{\sum_{12}} (\vec{y} - \vec{y}_i) \right)^2 \right], \qquad (11)
\end{aligned}$$

where $\mathrm{Dist}^{\mathrm{marginal}}$ is the distance of marginal distributions between the source and target domains. Also, D_S and D_T demonstrate the set of instances in the source and target domains, in turn. The discrepancy across the marginal distributions $P(X_{\mathrm{S}})$ and $P(X_T)$ is reduced by minimizing $\mathrm{Dist}^{\mathrm{marginal}}(D_{\mathrm{S}}, D_T)$.

Even though decreasing the marginal distribution difference across training and target domains minimizes the domain misalignment, but the conditional distribution difference of domains should be considered for a robust distribution adaptation. However, as the target domain lacks labels, conditional distribution adaptation is a nontrivial problem. We apply a trained classifier on the source samples to measure the posterior probabilities. Using this technique, the pseudo-labels of the target samples are obtained. We rewrite the Bregman divergence measure to match the class conditional distributions as follows:

$$\begin{aligned}
&\mathrm{Dist}^{\mathrm{conditional}} \textstyle\sum_{c=1}^{C} (D_{\mathrm{S}^c}, D_{T^c}) \\
&= \left[\int \left(\frac{1}{n_{\mathrm{s}}^c} \sum_{i=1}^{n_{\mathrm{s}}^c} \mathbf{G}_{\sum_1} (\vec{y} - \vec{y}_i) \right)^2 d\vec{y} \right] \\
&+ \left[\int \left(\frac{1}{n_{\mathrm{t}}^c} \sum_{j=n_{\mathrm{s}}^c+1}^{n_{\mathrm{s}}^c+n_{\mathrm{t}}^c} \mathbf{G}_{\sum_2} (\vec{y} - \vec{y}_i) \right)^2 d\vec{y} \right] \\
&- \left[\left(\frac{1}{n_{\mathrm{s}}^c n_{\mathrm{t}}^c} \sum_{i=1}^{n_{\mathrm{s}}^c} \sum_{j=n_{\mathrm{s}}^c+1}^{n_{\mathrm{s}}^c+n_{\mathrm{s}}^c} \mathbf{G}_{\sum_{12}} (\vec{y} - \vec{y}_i) \right)^2 \right], \qquad (12)
\end{aligned}$$

where $\mathrm{Dist}^{\mathrm{conditional}}$ is the class-conditional distributions distance across the source and target domains. Moreover, n_{s}^c and n_{t}^c denote the number of examples in source and target domains that belong to class c, respectively. Also, D_{S^c} is the set of examples belonging to class c in source data, and D_{T^c} is the set of examples belonging to class c in target data. With minimizing $\mathrm{Dist}^{\mathrm{conditional}} \sum_{c=1}^{C} (D_{\mathrm{S}^c}, D_{T^c})$, the conditional distribution mismatches between D_{S^c} and D_{T^c} are reduced.

3.3 Unsupervised domain adaptation via transferred local Fisher discriminant analysis (TLFDA)

The intuition behind TLFDA is to minimize the marginal and conditional distribution mismatches between the source and target domains by finding an optimal couple of projection and classification models. To this end, LFDA as a dimensionality reduction method is exploited to find a latent subspace with the criteria embedded in Eqs. 3, 11, and 12 as follows:

$$\begin{aligned}
J_{TLFDA} = \mathrm{argmin}_{J \in \mathbb{R}^{D \times d}} [& J_{\mathrm{LFDA}} \\
& + \lambda (\mathrm{Dist}^{\mathrm{marginal}}(D_{\mathrm{S}}, D_T) \\
& + \mathrm{Dist}^{\mathrm{conditional}} \textstyle\sum_{c=1}^{C} (D_{\mathrm{S}^c}, D_{T^c}))], \qquad (13)
\end{aligned}$$

where λ denotes the regularization parameter to balance between feature matching and domain adaptation. The first part of the equation is a transformation matrix that maps samples into a latent subspace. The second and third parts are minimizing the marginal and conditional distribution mismatches across domains, respectively. Equation 13 can be treated using the gradient descent algorithm, i.e.,

$$\begin{aligned}
J \leftarrow J - \eta (& \partial_J J_{\mathrm{LFDA}} + \lambda (\partial_J \mathrm{Dist}^{\mathrm{marginal}}(D_{\mathrm{S}}, D_T) \\
& + \partial_J \mathrm{Dist}^{\mathrm{conditional}} \textstyle\sum_{c=1}^{C} (D_{\mathrm{S}^c}, D_{T^c})), \qquad (14)
\end{aligned}$$

where η is the learning rate and ∂_J is the gradient with respect to J. The derivative of J_{LFDA} with respect to J is given by

$$\begin{aligned}
\frac{\partial_J J_{\mathrm{LFDA}}}{\partial_J} &= 2\mathrm{tr}^{-1} (J^{\mathrm{T}} \tilde{S}^{(b)} J)^{-1} \tilde{S}^{(w)} J \\
&\quad - 2\mathrm{tr}^{-2} (J^{\mathrm{T}} \tilde{S}^{(b)} J) tr (J^{\mathrm{T}} \tilde{S}^{(w)} J) \tilde{S}^{(b)} J. \qquad (15)
\end{aligned}$$

To obtain the optimal linear subspace J in Eq. 13, a direct method is to optimize Eq. 13 with respect to J iteratively by adopting the gradient descent technique as follows:

$$\begin{aligned}
J_{k+1} = J_k - \eta(k) \Bigg(& \frac{\partial_J J_{\mathrm{LFDA}}}{\partial_J} \\
& + \lambda \left(\sum_{i=1}^{n_s+n_t} \frac{\partial_J \mathrm{Dist}^{\mathrm{marginal}}(D_{\mathrm{S}}, D_T)}{\partial \vec{y}_i} \frac{\partial \vec{y}_i}{\partial_J} \right. \\
& \left. + \sum_{j=1}^{n_{\mathrm{s}}^c+n_{\mathrm{t}}^c} \frac{\mathrm{Dist}^{\mathrm{conditional}}(D_{\mathrm{S}^c}, D_{T^c})}{\partial \vec{y}_i} \frac{\partial \vec{y}_i}{\partial_J} \right) \Bigg) \qquad (16)
\end{aligned}$$

where $\eta(k)$ is the learning rate factor at kth iteration that controls the gradient step size. According to the quadratic form Eq. 11, the derivative of $\text{Dist}^{\text{marginal}}$ with respect to J is

$$\begin{aligned}
&\sum_{i=1}^{n_s+n_t} \frac{\partial_J \text{Dist}^{\text{marginal}}(D_S, D_T)}{\partial \vec{y}_i} \frac{\partial \vec{y}_i}{\partial J} \\
&= \sum_{i=1}^{n_s} \frac{\partial_J \text{Dist}^{\text{marginal}}(D_S, D_T)}{\partial \vec{y}_i} \vec{x}_i^T \\
&\quad + \sum_{i=1}^{n_s+n_t} \frac{\partial_J \text{Dist}^{\text{marginal}}(D_S, D_T)}{\partial \vec{y}_i} \vec{x}_i^T \\
&= \frac{1}{n_s^2} \sum_{s=1}^{n_s} \sum_{t=1}^{n_t} \mathbf{G}_{\sum_{11}}(\vec{y}_s - \vec{y}_t) \\
&\quad + \frac{1}{n_t^2} \sum_{s=1}^{n_s+n_t} \sum_{t=1}^{n_s+n_t} \mathbf{G}_{\sum_{22}}(\vec{y}_s - \vec{y}_t) \\
&\quad - \frac{1}{n_s n_t} \sum_{s=1}^{n_s} \sum_{t=n_s+1}^{n_s+n_t} \mathbf{G}_{\sum_{12}}(\vec{y}_s - \vec{y}_t). \qquad (17)
\end{aligned}$$

And according to the quadratic form Eq. 12, the derivative of $\text{Dist}^{\text{conditional}}$ with respect to J is

$$\begin{aligned}
&\sum_{i=1}^{n_s^c+n_t^c} \frac{\partial_J \text{Dist}^{\text{conditional}}(D_{S^c}, D_{T^c})}{\partial \vec{y}_i} \frac{\partial \vec{y}_i}{\partial J} \\
&= \sum_{i=1}^{n_s^c} \frac{\partial_J \text{Dist}^{\text{conditional}}(D_{S^c}, D_{T^c})}{\partial \vec{y}_i} \vec{x}_i^T \\
&\quad + \sum_{i=n_s^c+1}^{n_s^c+n_t^c} \frac{\partial_J \text{Dist}^{\text{conditional}}(D_{S^c}, D_{T^c})}{\partial \vec{y}_i} \vec{x}_i^T \\
&= \frac{1}{(n_s^c)^2} \sum_{s=1}^{n_s^c} \sum_{t=1}^{n_t^c} \mathbf{G}_{\sum_{11}}(\vec{y}_s - \vec{y}_t) \\
&\quad + \frac{1}{(n_t^c)^2} \sum_{s=n_s^c+1}^{n_s^c+n_t^c} \sum_{t=n_s^c+1}^{n_s^c+n_t^c} \mathbf{G}_{\sum_{22}}(\vec{y}_s - \vec{y}_t) \\
&\quad - \frac{1}{n_s^c n_t^c} \sum_{s=1}^{n_s^c} \sum_{t=n_s^c+1}^{n_s^c+n_t^c} \mathbf{G}_{\sum_{12}}(\vec{y}_s - \vec{y}_t). \qquad (18)
\end{aligned}$$

For two optional Gaussian kernels, we have $\int \mathbf{G}_{\sum_1}(\vec{y} - \vec{y}_s)\mathbf{G}_{\sum_2}(\vec{y} - \vec{y}_t)dy = \mathbf{G}_{\sum_1+\sum_2}(\vec{y}_s - \vec{y}_t)$ and, $\sum_{11} = \sum_1 + \sum_1$, $\sum_{22} = \sum_2 + \sum_2$ and $\sum_{12} = \sum_1 + \sum_2$. Based on Eqs. 15, 17, and 18, TLFDA is solved iteratively subject to $J^{\mathrm{T}} J = I$.

The computational complexity of TLFDA is investigated. We analyze the computational complexity of TLFDA using big O notation where n_s and n_t denote the number of source samples and the number of target samples, respectively. Moreover, n_s^c and n_t^c denote the number of instances in the source and target domains that belong to class c, respectively. The computational cost for Eqs. 15, 17, and 18 are $O((n_s + n_t)^2)$, $O((n_s + n_t)^2)$, and $O((n_s^c + n_t^c)^2)$, respectively. Hence, the total computational complexity of TLFDA is $O((n_s + n_t)^2)$.

4 Experimental setup

In this section, we introduce the domain adaptation benchmark datasets and the implementation details of TLFDA and other compared methods.

4.1 Data description

We apply our experiments on the following four visual domain adaptation benchmarks: Office+Caltech-256 (Surf) [41, 42], Office+Caltech-256 (Decaf_6) [43], Digit (USPS [44] and MNIST [45]), and CMU-PIE [46].

Office benchmark contains three domains with 31 object classes where domains either are downloaded from commercial sites (e.g. amazon.com) or taken with high-resolution digital SLR cameras or captured by low-resolution webcams. The set contains 4110 images which has 10 common classes with a minimum of 7 and a maximum of 100 examples in each class over three domains. Caltech-256 includes images in which objects appear in several various poses. Thus, the set contains images that are not normally aligned. The dataset contains 256 categories with a minimum of 80 and a maximum of 827 images in each category. We make 12 cross-domain tasks according to four Office+Caltech-256 (Surf) benchmarks via considering two different domains as the source and target domains. Also, we utilize Decaf_6 (deep convolutional activation feature) features with 4096 dimensions normalized to unit vectors. Decaf_6 features are the activation values of the 6th layer of a convolutional neural network (CNN) trained on the ImageNet dataset [43]. Though, we are to compare the effectiveness of TLFDA with traditional and other deep DA methods.

USPS (U) and MNIST (M) domains are popular handwritten digit benchmarks with various statistics and distributions. The USPS dataset possesses 7291 training and 2007 test images with overall, 9298 images with 16×16 size scanned from envelopes of the US Postal Service. MNIST dataset contains 60,000 training and 10,000 test images with 28×28 size scanned from mixed American Census Bureau employees and American high school students. All images of USPS and MNIST datasets are resized to 16×16 with a grayscale level. Therefore, we design two cross-domain tasks as follows: $U \longrightarrow M$ and $M \longrightarrow U$.

Table 1 Classification accuracy (%) of the proposed method on Office+Caltech-256 (Surf) and Digits datasets

Dataset	FDA	LPP	JACRL (2017)	RTML (2017)	VDA (2017)	JGSA (2017)	CLGA (2018)	DICD (2018)	DGA-DA (2020)	JDA-CDMA (2020)	TTLC (2021)	DOLL-DA (2021)	TLFDA
$C \longrightarrow A$	40.22	37.58	56.26	43.5	46.14	51.46	48.02	47.29	52.09	53.65	**56.68**	54.18	51.46
$C \longrightarrow W$	40.11	38.98	47.8	45.5	46.1	45.42	42.37	46.44	47.12	**54.58**	51.86	51.19	45.62
$C \longrightarrow D$	39.99	42.04	43.95	49.7	**51.59**	45.86	49.04	49.68	45.86	50.32	45.22	47.13	45.86
$A \longrightarrow C$	41.36	37.58	42.65	42.7	42.21	41.50	42.3	42.39	41.32	41.94	40.34	44.88	**47.59**
$A \longrightarrow W$	41.65	35.93	41.69	43.4	51.19	45.76	41.36	45.08	38.31	50.17	**55.25**	45.08	46.76
$A \longrightarrow D$	40.89	39.94	43.31	43.3	48.41	47.13	36.31	38.85	38.22	38.85	**57.32**	46.50	47.83
$W \longrightarrow C$	41	26.71	34.64	36.9	27.6	33.21	32.95	33.57	33.30	32.95	30.54	38.29	**43.39**
$W \longrightarrow A$	32.90	37.77	39.25	37.5	26.1	39.87	34.57	34.13	**41.75**	38.55	39.87	39.46	39.87
$W \longrightarrow D$	41.52	73.25	85.99	91.7	89.18	90.45	**92.36**	89.81	89.81	82.80	89.81	86.62	90.85
$D \longrightarrow C$	**43.21**	26.18	35.17	37.0	31.26	29.92	33.66	34.64	33.66	34.28	31.43	32.41	32.92
$D \longrightarrow A$	42.56	36.78	37.89	36.3	37.68	38	**89.83**	34.45	33.61	42.28	40.81	38.20	44.69
$D \longrightarrow W$	42.91	39.12	89.15	90.5	90.85	91.86	35.99	91.19	**93.22**	87.46	91.86	90.22	91.86
$U \longrightarrow M$	73.51	44.7	42.15	61.82	62.95	68.15	58.35	65.2	70.75	70.2	69.15	86.90	**89.54**
$M \longrightarrow U$	64.89	65.94	63.56	69.52	74.72	80.44	71.28	77.83	82.33	84.5	82.94	**97.20**	81.85
Average	44.76	41.6	49.96	52.09	51.86	53.50	50.60	52.18	52.95	54.47	55.93	57.02	**57.14**

Bold values represent the best statistically significant results

PIE is an introduced face benchmark containing 41,368 images with the size of 32×32 captured from 68 individuals. All images are taken by 13 synchronized cameras and 21 flashes with different poses, illuminations, and expressions. Depending on the position of images, the dataset is divided into five different subsets: PIE1(C05, left pose), PIE2(C07, upward pose), PIE3(C09, downward pose), PIE4(C27, front pose), and PIE5(C29, right pose). Hence, twenty cross-domain tasks are conducted as follows: $P1 \longrightarrow P2, P1 \longrightarrow P3, \ldots, P4 \longrightarrow P5$.

4.2 Method evaluation

The performance evaluation of our proposed approach is conducted on four DA datasets with two baseline machine learning methods (LPP and FDA) and ten state-of-the-art DA methods (JACRL [47], RTML [48], VDA [23], JGSA [49], CLGA [24], DICD [25], DGA-DA [26], JDA-CDMA [50], TTLC [27], and DOLL-DA [51]). Since TLFDA and other DA methods are dimensionality reduction techniques, we exploit a NN classifier to achieve classification results. Furthermore, TLFDA is compared with the best-reported results of the compared methods. Also, we evaluate the performance of TLFDA on Office+Caltech-256 (Decaf_6) dataset with a baseline deep method, AlexNet [52], and deep DA methods including DDC [53], AELM [54], and ELM [54]. Moreover, we evaluate TLFDA with DA methods including PUnDA [55], TAISL [56], SCA [57], and TIT [58] on Office+Caltech-256 (Decaf_6) benchmark.

4.3 Implementation details

To justly test and compare TLFDA with other methods, we measure the classification accuracy on the target domain (D_t) as the evaluation metric as follows:

$$\text{Accuracy} = \frac{|\ x :\in D_t \wedge f(x) = y(x)\ |}{n_t}, \tag{19}$$

where $f(x)$ is the achieved prediction function and $y(x)$ is the correct label of sample x, respectively. In addition, n_t is the number of target domain samples.

Furthermore, TLFDA consists of the following two free parameters. (1) λ is the regularization parameter in Eq. 13 that controls the trade-off between the feature matching and Bregman divergence, and (2) $\eta(k)$ is the learning rate factor at iteration k in Eq. 16, which controls the gradient step size for kth iteration. Additionally, the number of iterations to the convergence of TLFDA is set to 20.

5 Experimental results and discussion

In this section, the performance of TLFDA and other compared methods on a variety of visual DA benchmarks are compared.

5.1 Result evaluation

Tables 1 and 2 show the classification accuracies of TLFDA and other machine learning and DA approaches on object

Table 2 Classification accuracy (%) of the proposed method on PIE dataset

Dataset	FDA	LPP	JACRL (2017)	RTML (2017)	VDA (2017)	JGSA (2017)	CLGA (2018)	DICD (2018)	DGA-DA (2020)	JDA-CDMA (2020)	TTLC (2021)	DOLL-DA (2021)	TLFDA
$P1 \longrightarrow P2$	33.46	21.67	51.2	60.12	72.99	84.41	67.83	72.99	65.32	82.26	83.86	79.56	**92.88**
$P1 \longrightarrow P3$	33.58	23.59	57.6	55.21	61.64	69.98	63.85	72.00	62.81	77.88	**83.09**	78.59	78.75
$P1 \longrightarrow P4$	35.27	27.22	86.66	85.19	90.12	68.88	88.95	92.22	83.54	93.09	**96.37**	91.53	79.86
$P1 \longrightarrow P5$	34.78	15.50	52.39	52.98	42.40	57.35	61.76	66.85	56.07	57.05	77.21	75.94	**89.65**
$P2 \longrightarrow P1$	46.47	20.74	65.55	58.13	72.87	80.15	71.4	69.93	63.69	83.64	80.91	82.14	**86.95**
$P2 \longrightarrow P3$	43.29	37.62	68.5	63.92	75.61	89.58	72.98	65.87	61.27	82.60	80.39	75.61	**94.78**
$P2 \longrightarrow P4$	45.63	48.51	81.71	76.16	83.60	93.98	86.24	85.25	82.37	89.13	93.78	91.86	**95.98**
$P2 \longrightarrow P5$	35.43	22.24	54.53	40.38	57.72	93.43	51.23	48.71	46.63	71.57	77.88	72.90	**96.83**
$P3 \longrightarrow P1$	42.44	18.61	69.39	53.12	58.76	67.22	70.17	69.36	56.72	76.17	84.51	78.69	**87.82**
$P3 \longrightarrow P2$	69.62	36.53	61.76	58.67	74.65	87.50	73.48	65.44	61.26	78.39	84.9	78.59	**89.75**
$P3 \longrightarrow P4$	42.12	42.29	89.74	69.81	87.53	80.05	89.31	83.39	77.83	93.42	**97.48**	86.39	86.89
$P3 \longrightarrow P5$	37.27	21.26	60.54	42.13	52.63	65.16	55.51	61.4	44.24	76.84	**80.21**	72.12	79.76
$P4 \longrightarrow P1$	38.49	29.41	89.08	81.12	92.35	67.25	89.56	93.13	81.84	93.67	**98.56**	93.50	85.75
$P4 \longrightarrow P2$	39.29	55.25	85.64	83.92	92.27	86.95	92.94	90.12	85.27	95.58	**97.05**	90.61	88.95
$P4 \longrightarrow P3$	47.51	65.99	86.34	89.51	90.38	73.59	93.08	88.97	90.95	93.81	**94.06**	84.56	85.89
$P4 \longrightarrow P5$	34.11	31	76.04	56.26	69.98	71.10	71.63	75.61	53.80	84.93	**87.62**	86.97	82.86
$P5 \longrightarrow P1$	37.39	18.64	71.94	29.11	49.91	61.53	57.68	62.88	57.44	72.33	76.59	80.52	**86.93**
$P5 \longrightarrow P2$	39.62	22.65	47.45	33.28	62.31	92.41	55.43	57.03	53.84	76.37	79.07	86.32	**95.81**
$P5 \longrightarrow P3$	38.11	24.82	65.5	39.85	61.27	74.94	58.03	65.87	55.27	80.70	83.95	79.60	**86.84**
$P5 \longrightarrow P4$	49.08	28.96	79.9	47.13	71.19	74.28	71.85	74.77	61.82	84.20	**93.12**	83.90	83.98
Average	39.51	30.62	70.07	58.80	71	76.99	72.15	73.09	65.10	82.18	86.53	82.50	**87.85**

Bold values represent the best statistically significant results

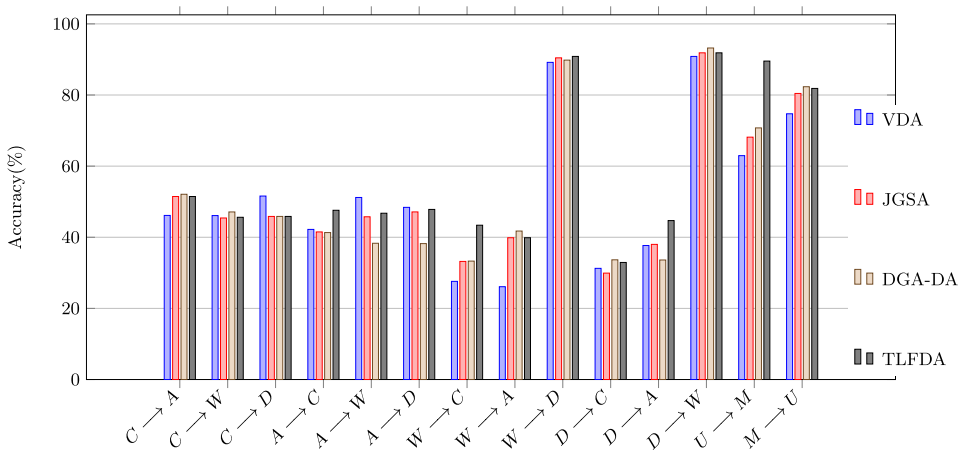

Fig. 1 Classification accuracy (%) on Office+Caltech-256 (Surf) and Digits datasets

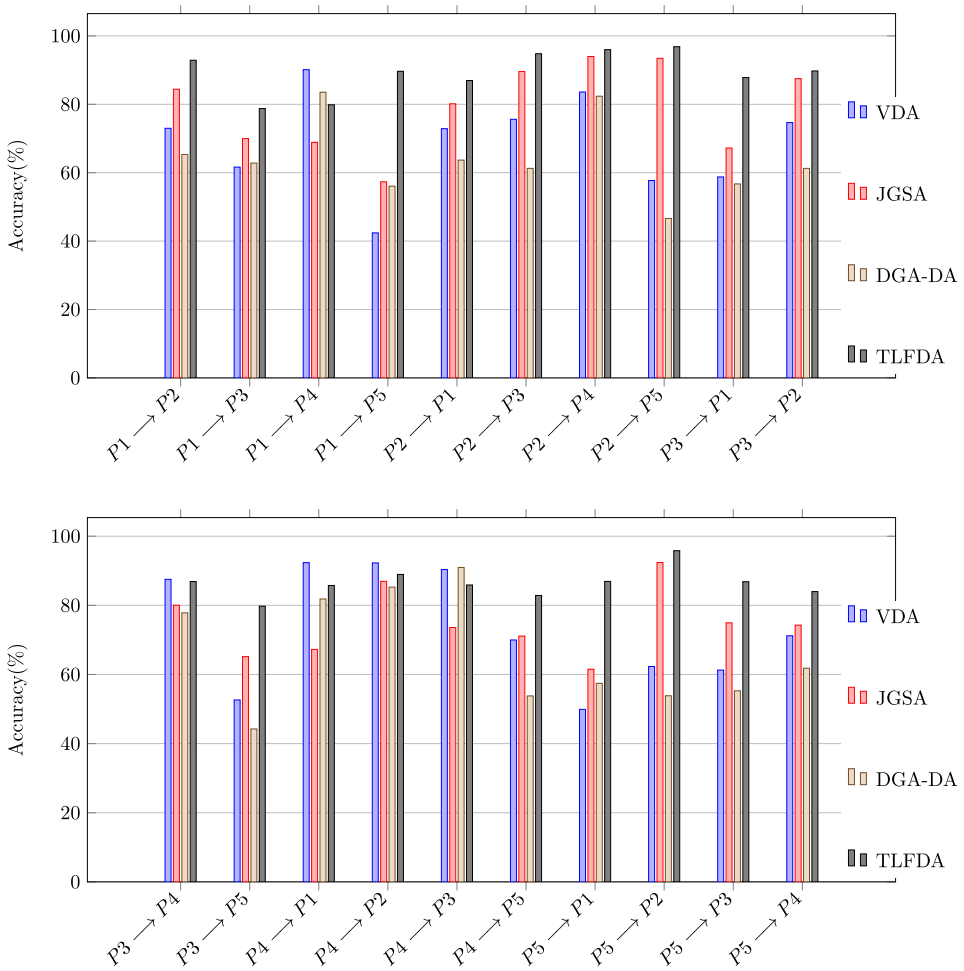

Fig. 2 Classification accuracy (%) on PIE datasets. **a** the first ten tasks, **b** the second ten tasks

recognition (Office+Caltech-256 (Surf)), hand-written Digits recognition (USPS, MNIST), and face (PIE) datasets, respectively. The results are visualized in Figs. 1 and 2 for better interpretation.

Due to the mismatched distribution among the training and test datasets, the performance improvement of TLFDA over FDA and LPP is (48.34%) and (57.23%) on the PIE dataset, respectively. In comparison to the best-compared approach TTLC on the PIE dataset, TLFDA achieves (1.32%) performance improvement. Moreover, TLFDA outperforms LPP and FDA in all tasks and in 11 tasks in comparison with TTLC on the PIE dataset. In the rest, our proposed method will be compared with other methods, in detail.

Table 3 Classification accuracy (%) of the proposed method on Office+Caltech-256 ($Decaf_6$) datasets

Dataset	AlexNet (2012)	DDC (2014)	AELM (2016)	ELM (2016)	PUnDA (2017)	TAISL (2017)	SCA (2017)	TIT (2018)	TLFDA
$C \longrightarrow A$	**91.9**	**91.9**	89.46	89.07	90.3	90	89.46	89.5	90.71
$C \longrightarrow W$	83.7	85.4	79.32	70.51	88.3	85.3	85.42	**92.1**	86.42
$C \longrightarrow D$	87.1	88.8	81.53	78.98	76.2	90.6	87.9	86.7	**91.69**
$A \longrightarrow C$	83	85	79.96	79.61	82.3	80.1	78.81	83.8	**88.77**
$A \longrightarrow W$	79.5	86.1	77.63	74.58	82.7	77.9	75.93	**91.4**	79.63
$A \longrightarrow D$	87.4	89	85.35	80.25	76.2	85.1	85.35	**89.1**	83.28
$W \longrightarrow C$	73	78	71.24	70.61	82.6	82.6	74.80	80.2	**85.32**
$W \longrightarrow A$	83.8	84.9	76.83	75.37	86.9	85.6	86.12	89.3	**89.66**
$W \longrightarrow D$	**100**	**100**	**100**	**100**	89.8	97.7	**100**	94.9	**100**
$D \longrightarrow C$	79	81.1	75.6	68.21	69.2	84	78.09	80.7	**87.36**
$D \longrightarrow A$	87.1	89.5	83.19	80.79	83.1	87.6	89.98	**92.5**	91.8
$D \longrightarrow W$	97.7	98.2	**98.98**	98.31	93.4	95.9	98.64	88.1	98.46
Average	86.1	88.2	83.25	80.52	83.42	86.9	85.88	88.2	**89.43**

Bold values represent the best statistically significant results

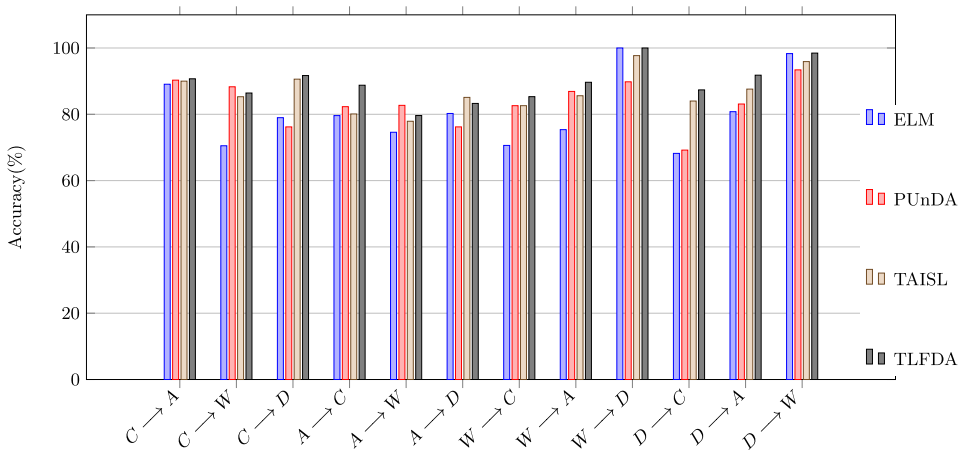

Fig. 3 Classification accuracy (%) on Office+Caltech-256 ($Decaf_6$) datasets

Most of the available methods that are based on FDA or LPP criteria do not achieve the desired results in case of domain shift problems where they do not consider the distribution mismatch between the source and target domains. FDA is a well-known approach that projects the data into a low-dimensional subspace with a linear combination of features. In dealing with shifted data, FDA can not generally transfer knowledge across domains.

LPP is another classical method of dimensionality reduction to find an embedding that preserves the local information. LPP uses a graph to model the geometrical structure of data. Despite the LPP efficiency, it cannot reduce the distribution divergence across the source and target domains due to the significant dissimilarity of scatters from the mean. Nevertheless, FDA shows better performance rather than LPP, because it uses discrimination in the feature extraction step. The results illustrate that TLFDA has (11.7%) and (13.07%) average classification accuracy improvement against FDA and LPP on Office+Caltech-256 (Surf) dataset, respectively. The performance improvement of TLFDA against FDA on the Digits dataset is (16.5%). Also, TLFDA obtains (30.38%) improvement compared to LPP on the Digits dataset.

JACRL is another state-of-the-art transfer learning method reducing the functional structural risk and the distribution mismatch across domains. Thus, JACRL learns an adaptive classifier through maximizing the manifold consistency of the adaptive classifier. However, TLFDA outperforms JACRL in most cases where it considers both the discriminative information contained in the training samples and the distribution bias between the training and test sets. TLFDA gains (7.18%) and (17.78%) performance improvement compared to JACRL on object+digit and PIE datasets, respectively.

RTML exploits the knowledge transfer to alleviate the domain shift in two directions, i.e., sample and feature space. RTML aims to build a cross-domain metric to reduce the mismatches across domains. However, TLFDA jointly benefits from representation and classification learning to adapt the

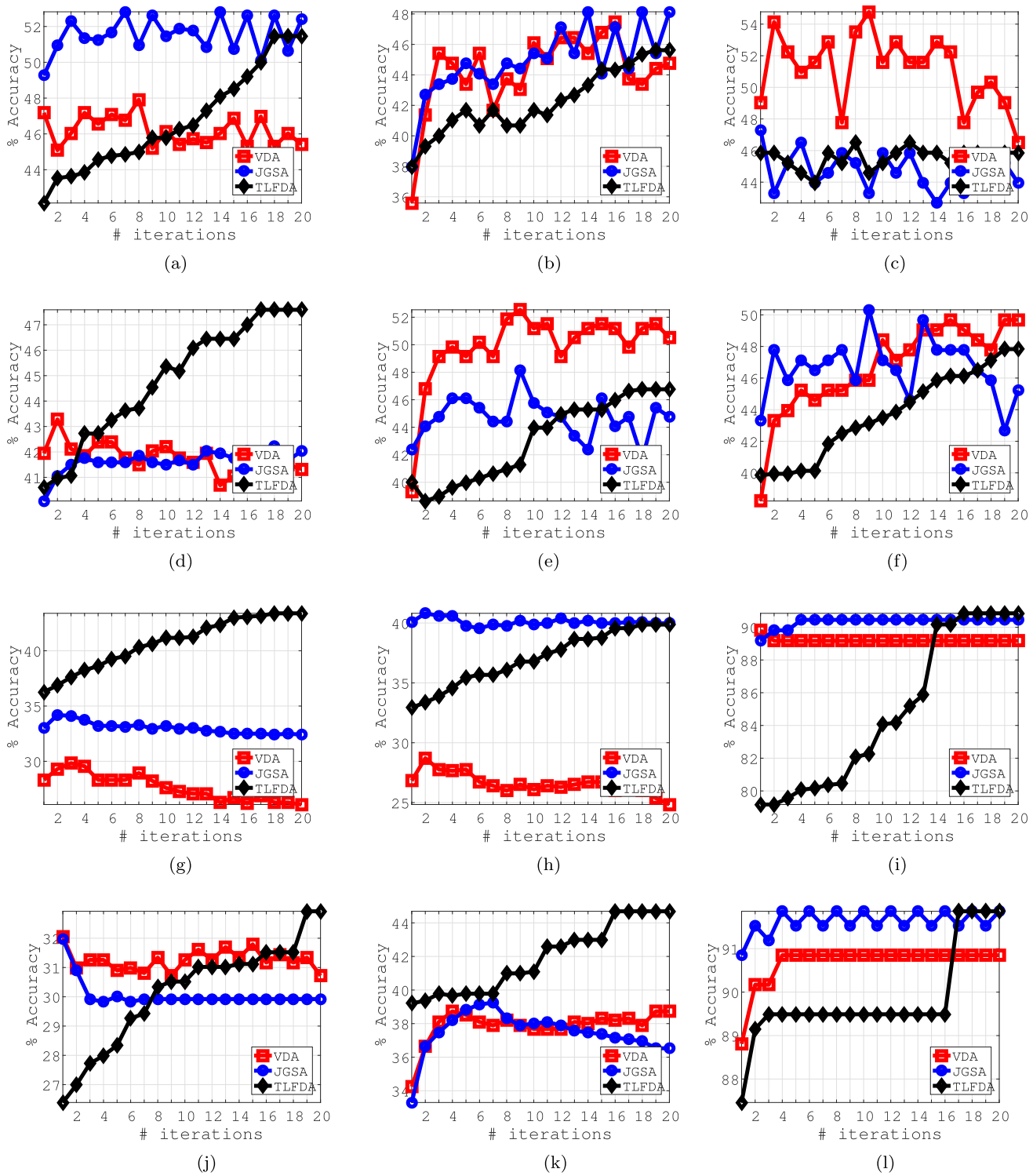

Fig. 4 Classification accuracy (%) with respect to the number of iterations for Office+Caltech-256 (Surf) dataset. **a** $C \longrightarrow A$. **b** $C \longrightarrow W$. **c** $C \longrightarrow D$. **d** $A \longrightarrow C$. **e** $A \longrightarrow W$. **f** $A \longrightarrow D$. **g** $W \longrightarrow C$. **h** $W \longrightarrow A$. **i** $W \longrightarrow D$. **j** $D \longrightarrow C$. **k** $D \longrightarrow A$. **l** $D \longrightarrow W$

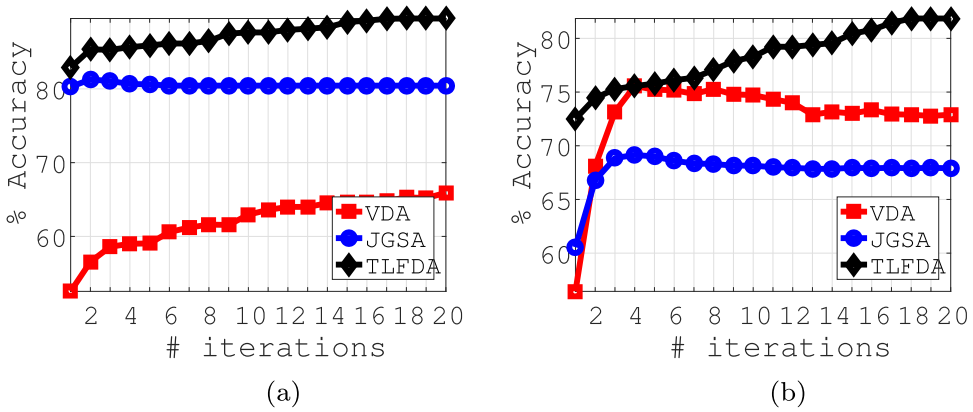

Fig. 5 Classification accuracy (%) with respect to the number of iterations for Digits dataset. **a** USPS versus MNIST. **b** MNIST versus USPS

source and target domains. TLFDA uses the source domain labels to construct the shared low-dimensional subspace and discriminate across various classes. TLFDA achieves (5.05%) and (29.05%) performance improvement in average classification accuracy compared to RTML on object+digit and PIE datasets, respectively.

VDA is a novel technique that exploits joint DA and TL to create a shared feature space that is robust against distribution mismatch. VDA discriminates different classes in the latent subspace by employing the domain invariant clustering technique. VDA only seeks to align the marginal and conditional distributions across the source and target domains, while it ignores the discriminative properties between different classes in the adapted domain. However, TLFDA minimizes the distances between the marginal and the conditional distributions of domains, while the specific information of domains (i.e., data manifold structure and within-class local structure) is preserved. TLFDA achieves (5.28%) and (16.85%) performance improvement in average classification accuracy compared to VDA on object+digit and PIE datasets, respectively.

JGSA proposes a unified framework to minimize the shift between domains both geometrically and statistically using both shared and domain-specific features of domains. JGSA aligns the source and target domains even with high divergence. However, the joint marginal and conditional distributions alignment between domains does not explicitly render the data discrimination in achieved feature representation. TLFDA gains (3.64%) and (10.86%) performance improvement compared to JGSA on object+digit and face datasets, respectively.

CLGA finds a unified subspace where the marginal and conditional distributions are globally matched. CLGA locally adapts both domains using the class and domain manifold structures. CLGA measures the distance across distributions via MMD whereas TLFDA employs Bregman divergence as a measurement metric. Moreover, TLFDA uses Bregman divergence to preserve the discrimination ability. TLFDA outperforms CLGA in 10 tasks out of 14 domain adaptation tasks and gains (6.54%) improvement on object+digit datasets. Also, TLFDA works better than CLGA with (15.7%) improvement on the face dataset.

DICD in a shared subspace adapts both the marginal and conditional distribution disparities. DICD discriminates classes by minimizing the distances across sample pairs in the same classes either in both domains. DICD maximizes the distances between samples with various class labels in both domains. However, TLFDA discriminates classes by preserving the intra-class and inter-class scatters through Bregman divergence. TLFDA outperforms DICD with (4.96%) and (14.76%) performance improvement on object+digit and face datasets, respectively.

DGA-DA, as a novel domain adaptation method, aligns the marginal and conditional distributions in a unified subspace using a non-parametric measurement method. Despite DGA-DA, TLFDA measures the distribution disparities across domains via Bregman divergence. TLFDA preserves the discriminated information of data via Bregman divergence. However, DGA-DA defines a repulsive form term to discriminate different classes. TLFDA works better than DGA-DA with (4.19%) accuracy improvement on object+digit datasets. Also, TLFDA outperforms DGA-DA in 18 tasks out of 20 tasks on the face dataset with (22.75%) accuracy improvement.

JDA-CDMA proposes a novel measurement metric called cross domain mean approximation (CDMA) to estimate the distances between the source and target samples. Then, based on CDMA, JDA-CDMA reduces the marginal and conditional divergences across both domains in a shared subspace. Although CDMA has low computational complexity in comparison with Bregman divergence, but TLFDA in comparison with JDA-CDMA gains (2.67%) and (5.67%) improvements on object+digit and face datasets, respectively.

TTLC aligns the marginal and conditional distributions in the respective subspaces via MMD. However, TLFDA adapts the marginal and conditional distribution disparities in a unified subspace via Bregman divergence. TTLC discriminates classes by creating condensed clusters in both domains. To

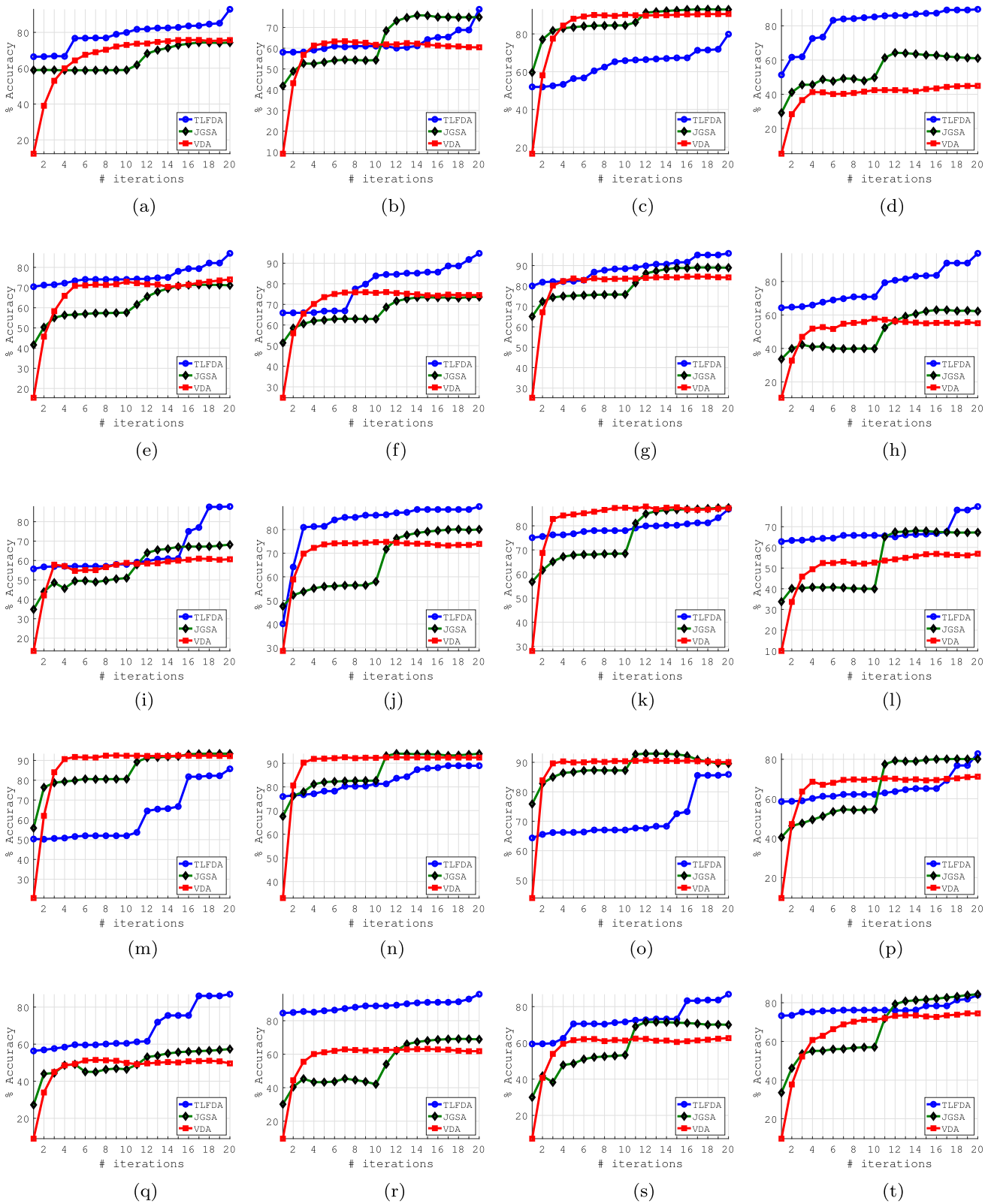

Fig. 6 Classification accuracy (%) with respect to the number of iterations for PIE dataset. **a** $P1 \longrightarrow P2$. **b** $P1 \longrightarrow P3$. **c** $P1 \longrightarrow P4$. **d** $P1 \longrightarrow P5$. **e** $P2 \longrightarrow P1$. **f** $P2 \longrightarrow P3$. **g** $P2 \longrightarrow P4$. **h** $P2 \longrightarrow P5$. **i** $P3 \longrightarrow P1$. **j** $P3 \longrightarrow P2$. **k** $P3 \longrightarrow P4$. **l** $P3 \longrightarrow P5$. **m** $P4 \longrightarrow P1$. **n** $P4 \longrightarrow P2$. **o** $P4 \longrightarrow P3$. **p** $P4 \longrightarrow P5$. **q** $P5 \longrightarrow P1$. **r** $P5 \longrightarrow P2$. **s** $P5 \longrightarrow P3$. **t** $P5 \longrightarrow P4$

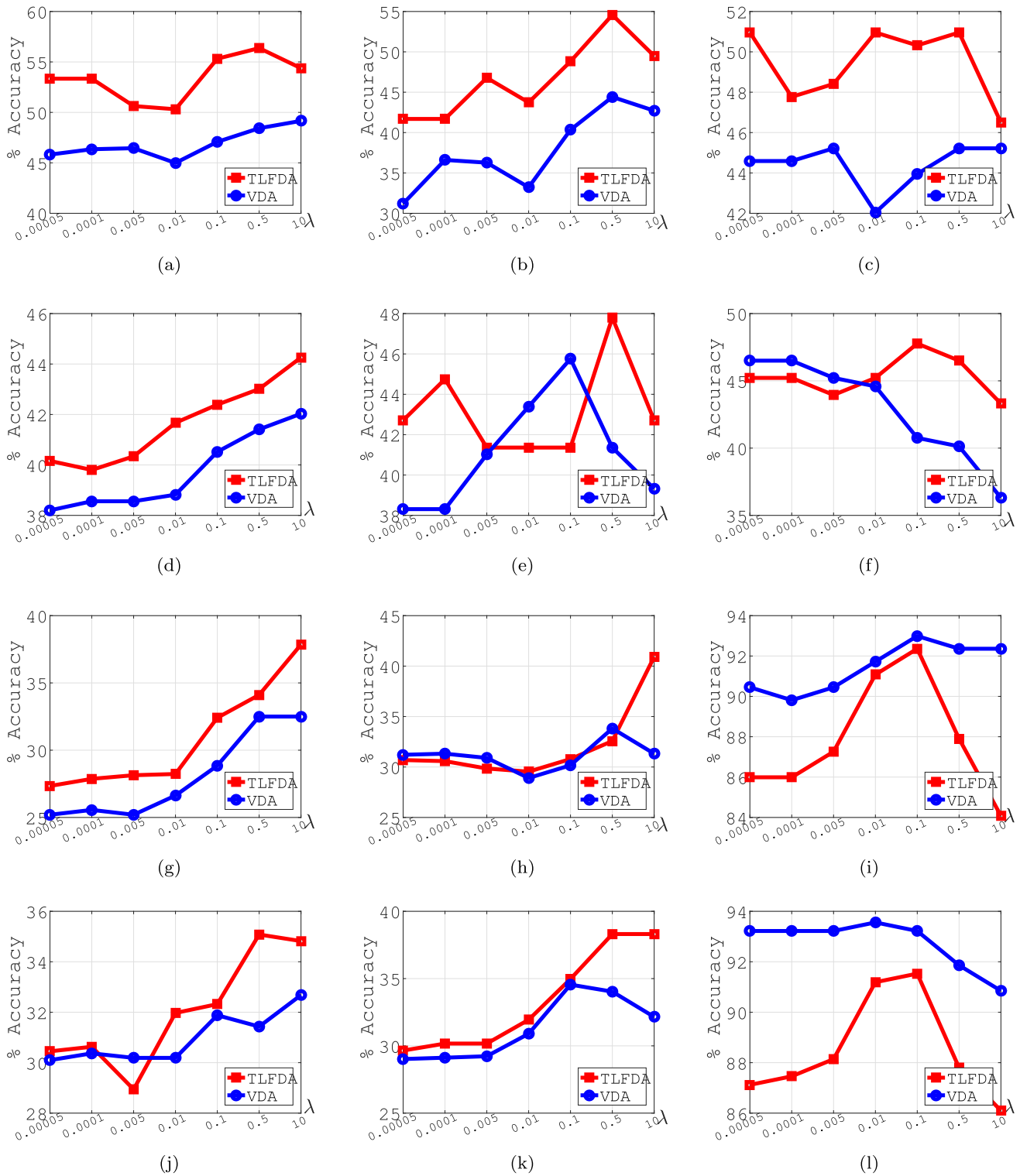

Fig. 7 Parameter evaluation with respect to classification accuracy (%) and parameter, λ, for Office+Caltech-256 (Surf) dataset. TLFDA obtains considerable results with large values of λ. We consider $\lambda = 1$ for Office+Caltech-256 (Surf) dataset. **a** $C \longrightarrow A$. **b** $C \longrightarrow W$. **c** $C \longrightarrow D$. **d** $A \longrightarrow C$. **e** $A \longrightarrow W$. **f** $A \longrightarrow D$. **g** $W \longrightarrow C$. **h** $W \longrightarrow A$. **i** $W \longrightarrow D$. **j** $D \longrightarrow C$. **k** $D \longrightarrow A$. **l** $D \longrightarrow W$

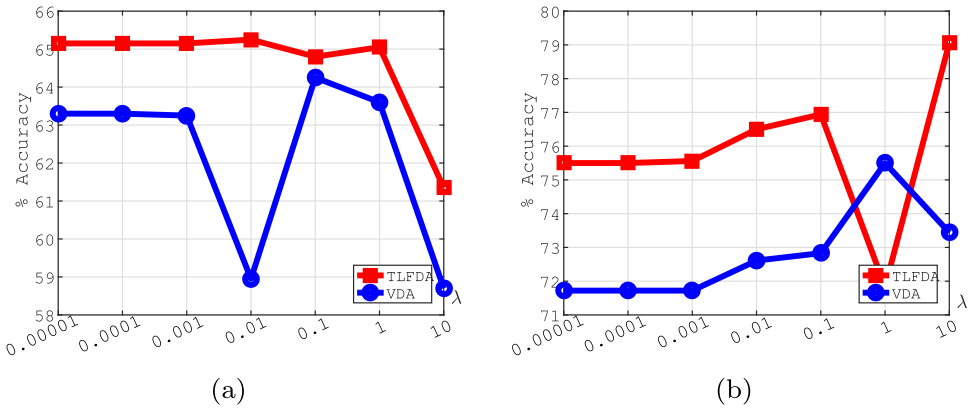

Fig. 8 Parameter evaluation with respect to the classification accuracy (%) and the regularization parameter, λ, for Digits dataset. TLFDA performs well on Digits dataset with small values of λ. We adjust $\lambda = 0.01$ on Digits dataset. **a** USPS versus MNIST. **b** MNIST versus USPS

this end, TTLC minimizes the distances across sample pairs in the same classes of both domains. In addition, TTLC maximizes the distances between each instance pairs of various classes in both domains. TLFDA gains (1.21%) and (1.32%) accuracy improvements on object+digit and face datasets, respectively.

DOLL-DA projects both domains into a common subspace by decreasing the marginal and conditional distribution discrepancies through adding the repulsive force to the MMD constraint. DOLL-DA uses a label embedding trick to propose an orthogonal label subspace. DOLL-DA proposes a noise-robust sparse orthogonal label regression term to prevent negative transfer learning and overfitting. However, TLFDA benefits from Bregman divergence as a measurement method. TLFDA outperforms DOLL-DA in 11 tasks out of 14 tasks on object+digit datasets. Also, TLFDA improves against DOLL-DA by achieving (5.35%) performance improvement on the face benchmark.

Figures 1 and 2 show the results evaluation of TLFDA comparing to DA methods including VDA, JGSA, and DGA-DA on Office+Caltech-256 (Surf) and Digit benchmarks with 14 tasks and on PIE dataset with 20 tasks, respectively. TLFDA outperforms VDA in 10 tasks out of 14 tasks and 15 tasks out of 20 tasks on PIE dataset. Figure 1 shows that TLFDA performs better than JGSA in 9 tasks. Figure 2 presents that TLFDA outperforms JGSA in all tasks on PIE dataset. Moreover, TLFDA has better accuracy in 7 tasks on Office+Caltech-256 (Surf) and Digit datasets and 18 tasks on PIE dataset.

In recent years, deep DA approaches have gained high performance. To compare the effectiveness of TLFDA with deep methods, we train TLFDA on Office+Caltech-256 (Decaf_6) datasets. Experimental results are shown in Table 3. According to the results, TLFDA outperforms deep methods including AlexNet, DDC, AELM, ELM in most of cross-domain tasks. TLFDA works better than AlexNet and DDC with (3.33%) and (1.23%) improvements, respectively. TLFDA outperforms ELM, PUnDA, and TAISL methods in most cases, where the results are visualized in Fig. 3. To be precise, Fig. 3 specifies that TLFDA is better than ELM and PUnDA in 10 tasks and TLFDA outperforms TAISL in 11 tasks of Office+Caltech-256 (Decaf_6) dataset. Moreover, TLFDA outperforms SCA and TIT as the domain adaptation methods on Office+Caltech-256 (Decaf_6) with (3.55%) and (1.23%), respectively. Although deep methods gain outperforming performances, they need to be trained on large amounts of data for reliable prediction. However, TLFDA outperforms deep methods while is trained on reasonable number of samples. Deep methods have large time complexities and need high-power processing equipment including GPU and CPU. However, TLFDA could be run on a medium-power CPU. Simplicity in processing and reliable predictions on enough number of samples cause that TLFDA to be picked in comparison to deep DA methods.

As the main findings of this study, TLFDA decreases the marginal and conditional distribution discrepancies via Bregman divergence in the mapped subspace. Moreover, TLFDA iteratively predicts pseudo-labels of the target domain via a model trained on the source domain.

5.2 Effectiveness evaluation

We conduct a variety of experiments in 20 iterations to evaluate the efficiency property of TLFDA. We run TLFDA, JGSA, and VDA in 20 iterations on Office+Caltech-256 (Surf), Digit, and PIE datasets. Figures 4, 5, and 6 illustrate the performance of TLFDA and two baseline methods with respect to the number of iterations on different benchmarks. As it is understood from the figures, in all datasets, TLFDA outperforms the best baseline method JGSA. Our proposed approach significantly reduces the difference of marginal and conditional distributions among the source and target domains. TLFDA predicts the accurate labels for target samples in an iterative manner. Almost, the predicted labels of each stage are better than the previous one.

The convergence property of TLFDA is evaluated in 20 iterations and its results are compared against JGSA and VDA. We run TLFDA on Office+Caltech-256 (Surf), digit,

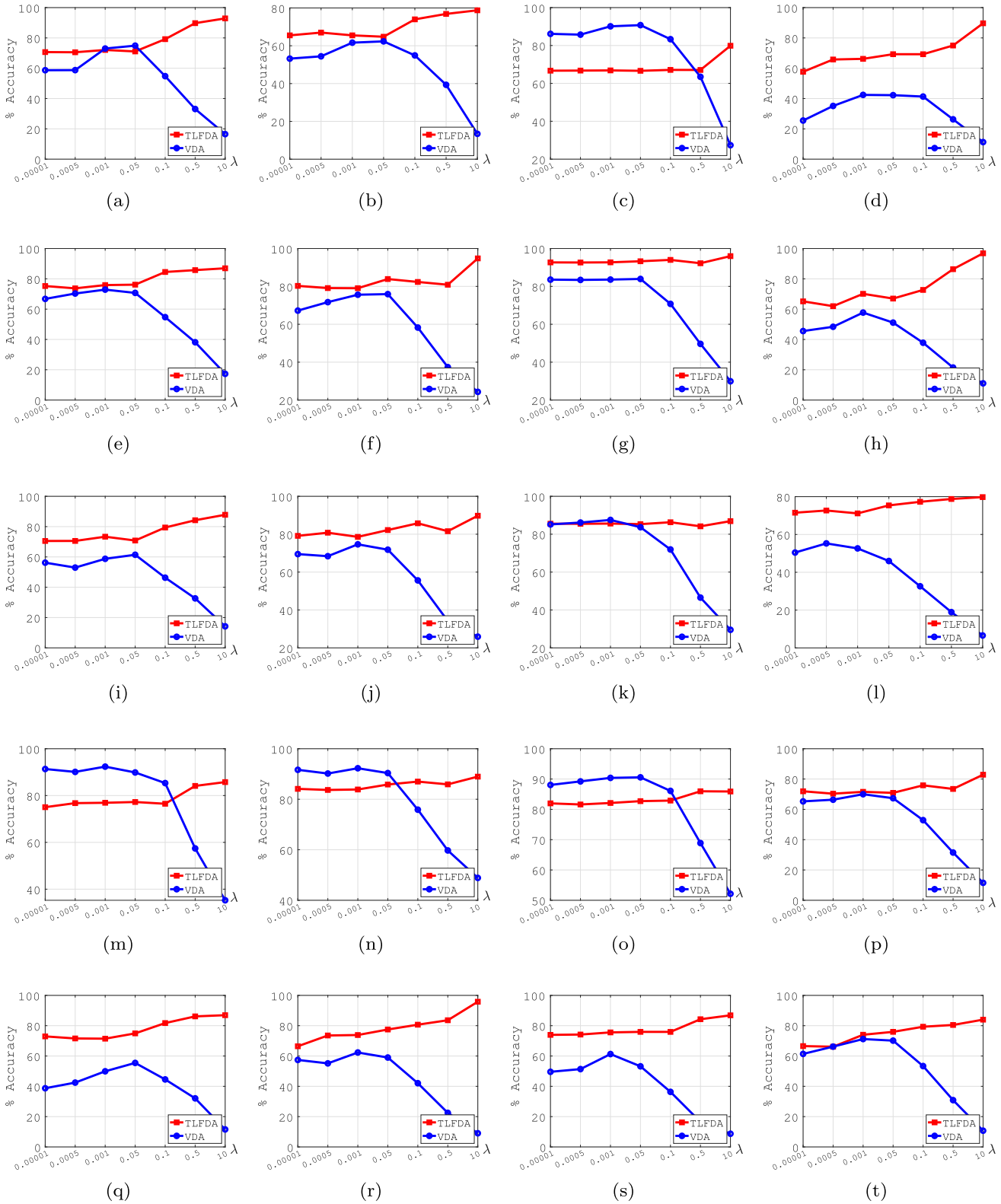

Fig. 9 Parameter evaluation with respect to the classification accuracy (%) and parameter, λ for PIE dataset. In most cases, TLFDA has better performance with $\lambda \in [0.00001\ \ 1]$. The optimal value of λ is 0.5 for PIE dataset. **a** $P1 \longrightarrow P2$. **b** $P1 \longrightarrow P3$. **c** $P1 \longrightarrow P4$. **d** $P1 \longrightarrow P5$. **e** $P2 \longrightarrow P1$. **f** $P2 \longrightarrow P3$. **g** $P2 \longrightarrow P4$. **h** $P2 \longrightarrow P5$. **i** $P3 \longrightarrow P1$. **j** $P3 \longrightarrow P2$. **k** $P3 \longrightarrow P4$. **l** $P3 \longrightarrow P5$. **m** $P4 \longrightarrow P1$. **n** $P4 \longrightarrow P2$. **o** $P4 \longrightarrow P3$. **p** $P4 \longrightarrow P5$. **q** $P5 \longrightarrow P1$. **r** $P5 \longrightarrow P2$. **s** $P5 \longrightarrow P3$. **t** $P5 \longrightarrow P4$

Table 4 Ablation study of 3 variants of TLFDA

Config methods	$TFLDA_{M+C}$	$TLFDA_M$	$TLFDA_C$	TLFDA
Average Office+Caltech-256 (Surf) Dataset	44.64	49.12	50.38	**52.39**
Average PIE Dataset	46.32	55.59	64.11	**87.85**
Average Digit Dataset	61.29	68.48	69.83	**85.7**

Bold values indicate the best results

and face datasets in 20 iterations and show the results in Figs. 4, 5, and 6, respectively. As is clear from the figures, TLFDA converges in 20 iterations in most cases. Although TLFDA fluctuates in some cases, it has a limited interlude after 20 iterations, and increasing the number of iterations does not have much effect on the performance improvement of the algorithm.

5.3 Parameter and ablation study

In Fig. 7, the classification accuracies of TLFDA and baseline methods are shown with respect to parameter λ on Office+Caltech-256 (Surf) dataset. $\lambda = 1$ is chosen for Office+Caltech-256 (Surf) dataset. Figure 8 shows the parameter evaluation with respect to the classification accuracy and parameter $\lambda \in [0.00001 \quad 10]$ for the Digits dataset. The reported results demonstrate that TLFDA operates well on the Digits dataset with small values of λ. Figure 9 illustrates the experimental results for parameter $\lambda \in [0.00001 \; 10]$ on the PIE dataset. As is obvious from the subfigures, TLFDA has better results with $\lambda = 0.5$ in most cases.

The performance of TLFDA is evaluated regarding the different values of parameters in various situations. To understand our model deeply, we evaluate several variants, i.e., (1) $TLFDA_M$ by eliminating the conditional distribution adaptation, (2) $TLFDA_C$ by eliminating the marginal distribution adaptation, and (3) $TLFDA_{M+C}$ by removing the marginal and conditional distributions adaptation, jointly. The evaluation results on various cases are shown in Table 4. The results indicate that $TLFDA_{M+C}$ performs worse than the other two variants whereas $TLFDA_C$ works better than others, since minimizing the diversity across the conditional distributions is crucial for robust distribution adaptation. However, all three variants cannot achieve better results than TLFDA. In fact, TLFDA constructs an effective feature representation for significant distribution misalignment across domains and it jointly aligns both the marginal and conditional distributions.

6 Conclusion and future works

In this paper, unsupervised domain adaptation via transferred local Fisher discriminant analysis (TLFDA) is proposed to deal with the shift and bias data of cross-domain problems. In TLFDA, a projection matrix is utilized to map the source and target domains into a shared subspace. Moreover, TLFDA reduces the joint marginal and conditional distribution mismatches based on Bregman divergence minimization. Experimental results on different visual benchmarks illustrate that TLFDA achieves better performance where there always exist shifts and biased data across domains. However, TLFDA has not been investigated to deal with big data. In the future, we will plan to merge transfer and deep learning for big data problems, which enables the deep learning network to transfer across cross-distribution sets.

Author contributions The methodology, writing and theoretical analysis: MZ. Supervision: JT. Writing and analysis: SR.

Funding No funding was received for the submitted work.

Data availability Datasets are available and from previous works [41–46].

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

Ethics approval Not applicable.

References

1. Ahmadvand, M., Tahmoresnezhad, J.: Metric transfer learning via geometric knowledge embedding. Appl. Intell. **51**(2), 921–934 (2021)
2. Shao, L., Zhu, F., Li, X.: Transfer learning for visual categorization: a survey. IEEE Trans. Neural Netw. Learn. Syst. **26**(5), 1019–1034 (2014)
3. Li, J., Yue, W., Ke, L.: Structured domain adaptation. IEEE Trans. Circ. Syst. Video Technol. **27**(8), 1700–1713 (2016)
4. Tahmoresnezhad, J., Hashemi, S.: Diret: an effective discriminative dimensionality reduction approach for multi-source transfer learning. Sci. Iran. **24**(3), 1303–1311 (2017)
5. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
6. He, X., Yan, S., Yuxiao, H., Niyogi, P., Zhang, H.-J.: Face recognition using laplacianfaces. IEEE Trans. Pattern Anal. Mach. Intell. **27**(3), 328–340 (2005)
7. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
8. Wand, M.P., Jones, M.C.: Kernel Smoothing. CRC Press, London (1994)

9. Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. J. Mach. Learn. Res. **8**(5), 1 (2007)
10. Tian, L., Tang, Y., Hu, L., Ren, Z., Zhang, W.: Domain adaptation by class centroid matching and local manifold self-learning. IEEE Trans. Image Process. **29**, 9703–9718 (2020)
11. Noori Saray, S., Tahmoresnezhad, J.: Joint distinct subspace learning and unsupervised transfer classification for visual domain adaptation. Signal Image Video Process. **15**(2), 279–287 (2021)
12. Sun, Q., Chattopadhyay, R., Panchanathan, S., Ye, J.: A two-stage weighting framework for multi-source domain adaptation. Adv. Neural Inf. Process. Syst. **24**, 1 (2011)
13. Ishii, M., Sato, A.: Joint optimization of feature transform and instance weighting for domain adaptation. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 3793–3799. IEEE, New York(2017)
14. Gong, B., Grauman, K., Sha, F.: Connecting the dots with landmarks: discriminatively learning domain-invariant features for unsupervised domain adaptation. In: International Conference on Machine Learning, pp. 222–230. PMLR (2013)
15. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. Adv. Neural Inf. Process. Syst. **19**, 1 (2006)
16. Aljundi, R., Emonet, R., Muselet, D., Sebban, M.: Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 56–63 (2015)
17. Aytar, Y., Zisserman, A.: Tabula rasa: model transfer for object category detection. In: 2011 International Conference on Computer Vision, pp. 2252–2259. IEEE (2011)
18. Jiang, W., Zavesky, E., Chang, S.-F., Loui, A.: Cross-domain learning methods for high-level visual concept classification. In: 2008 15th IEEE International Conference on Image Processing, pp. 161–164. IEEE (2008)
19. Duan, L., Tsang, I.W., Xu, D., Chua, T.-S.: Domain adaptation from multiple sources via auxiliary classifiers. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 289–296 (2009)
20. Long, M., Wang, J., Ding, G., Pan, S.J., Philip, S.Y.: Adaptation regularization: a general framework for transfer learning. IEEE Trans. Knowl. Data Eng. **26**(5), 1076–1089 (2013)
21. Long, M., Wang, J., Ding, G., Sun, J., Philip, S.Y.: Transfer feature learning with joint distribution adaptation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2200–2207 (2013)
22. Yong, X., Fang, X., Jian, W., Li, X., Zhang, D.: Discriminative transfer subspace learning via low-rank and sparse representation. IEEE Trans. Image Process. **25**(2), 850–863 (2015)
23. Tahmoresnezhad, J., Hashemi, S.: Visual domain adaptation via transfer feature learning. Knowl. Inf. Syst. **50**(2), 585–605 (2017)
24. Liu, J., Li, J., Ke, L.: Coupled local-global adaptation for multi-source transfer learning. Neurocomputing **275**, 247–254 (2018)
25. Li, S., Song, S., Huang, G., Ding, Z., Cheng, W.: Domain invariant and class discriminative feature learning for visual domain adaptation. IEEE Trans. Image Process. **27**(9), 4260–4273 (2018)
26. Luo, L., Chen, L., Shiqiang, H., Ying, L., Wang, X.: Discriminative and geometry-aware unsupervised domain adaptation. IEEE Trans. Cybern. **50**(9), 3914–3927 (2020)
27. Rezaei, S., Tahmoresnezhad, J., Solouk, V.: A transductive transfer learning approach for image classification. Int. J. Mach. Learn. Cybern. **12**(3), 747–762 (2021)
28. Sun, J., Wang, Z., Wang, W., Li, H., Sun, F.: Domain adaptation with geometrical preservation and distribution alignment. Neurocomputing **454**, 152–167 (2021)
29. Liu, W., Li, J., Liu, B., Guan, W., Zhou, Y., Changsheng, X.: Unified cross-domain classification via geometric and statistical adaptations. Pattern Recogn. **110**, 107658 (2021)
30. Sanodiya, R.K., Paul, D., Yao, L., Mathew, J., Juhi, A.: A feature selection approach to visual domain adaptation in classification. In: International Conference on Neural Information Processing, pp 77–89. Springer, London (2020)
31. Luo, L., Wang, X., Hu, S., Wang, C., Tang, Y., Chen, L.: Close yet distinctive domain adaptation. Preprint arXiv:1704.04235 (2017)
32. Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. Adv. Neural Inf. Process. Syst. **13**, 1 (2000)
33. Kennedy, J.: Particle swarm optimization. In: Encyclopedia of Machine Learning, pp. 760–766. Springer, Berlin (2011)
34. Wang, J., Feng, W., Chen, Y., Yu, H., Huang, M., Philip, S.Y.: Visual domain adaptation with manifold embedded distribution alignment. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 402–410 (2018)
35. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5018–5027 (2017)
36. Minnehan, B., Savakis, A.: Deep domain adaptation with manifold aligned label transfer. Mach. Vis. Appl. **30**(3), 473–485 (2019)
37. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
38. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. Preprint arXiv:1603.04779 (2016)
39. De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L.: The mahalanobis distance. Chemomet. Intell. Lab. Syst. **50**(1), 1–18 (2000)
40. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
41. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: European Conference on Computer Vision, pp. 213–226. Springer, London (2010)
42. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007)
43. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: a deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning, pp. 647–655. PMLR (2014)
44. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
45. Hull, J.J.: A database for handwritten text recognition research. IEEE Trans. Pattern Anal. Mach. Intell. **16**(5), 550–554 (1994)
46. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: Proceedings of fifth IEEE International Conference on Automatic Face Gesture Recognition, pp. 53–58. IEEE, New York (2002)
47. Gheisari, M., Baghshah, M.S.: Joint predictive model and representation learning for visual domain adaptation. Eng. Appl. Artif. Intell. **58**, 157–170 (2017)
48. Ding, Z., Yun, F.: Robust transfer metric learning for image classification. IEEE Trans. Image Process. **26**(2), 660–670 (2016)
49. Zhang, J., Li, W., Ogunbona, P.: Joint geometrical and statistical alignment for visual domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1859–1867 (2017)
50. Zang, S., Cheng, Y., Wang, X., Qiang, Yu., Xie, G.-S.: Cross domain mean approximation for unsupervised domain adaptation. IEEE Access **8**, 139052–139069 (2020)

51. Luo, L., Chen, L., Hu, S.: Discriminative noise robust sparse orthogonal label regression-based domain adaptation. Preprint arXiv:2101.04563 (2021)
52. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**, 1 (2012)
53. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: maximizing for domain invariance. Preprint arXiv:1412.3474 (2014)
54. Uzair, M., Mian, A.: Blind domain adaptation with augmented extreme learning machine features. IEEE Trans. Cybern. **47**(3), 651–660 (2016)
55. Gholami, B., Pavlovic, V., et al.: Punda: probabilistic unsupervised domain adaptation for knowledge transfer across visual categories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3581–3590 (2017)
56. Lu, H., Zhang, L., Cao, Z., Wei, W., Xian, K., Shen, C., van den Hengel, A.: When unsupervised domain adaptation meets tensor representations. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 599–608 (2017)
57. Ghifary, M., Balduzzi, D., Kleijn, W.B., Zhang, M.: Scatter component analysis: a unified framework for domain adaptation and domain generalization. IEEE Trans. Pattern Anal. Mach. Intell. **39**(7), 1414–1430 (2016)
58. Li, J., Lu, K., Huang, Z., Zhu, L., Shen, H.T.: Transfer independently together: a generalized framework for domain adaptation. IEEE Trans. Cybern. **49**(6), 2144–2155 (2018)