The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Domain-Shared Group-Sparse Dictionary Learning for Unsupervised Domain Adaptation

Baoyao Yang,[1] **Andy J. Ma,**[1,2] **Pong C. Yuen**[1]
[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[2]School of Data and Computer Science, Sun Yat-sen University, China
byyang@comp.hkbu.edu.hk, majh8@mail.sysu.edu.cn, pcyuen@comp.hkbu.edu.hk

Abstract

Unsupervised domain adaptation has been proved to be a promising approach to solve the problem of dataset bias. To employ source labels in the target domain, it is required to align the joint distributions of source and target data. To do this, the key research problem is to align conditional distributions across domains without target labels. In this paper, we propose a new criterion of domain-shared group-sparsity that is an equivalent condition for conditional distribution alignment. To solve the problem in joint distribution alignment, a domain-shared group-sparse dictionary learning method is developed towards joint alignment of conditional and marginal distributions. A classifier for target domain is trained using the domain-shared group-sparse coefficients and the target-specific information from the target data. Experimental results on cross-domain face and object recognition show that the proposed method outperforms eight state-of-the-art unsupervised domain adaptation algorithms.

1 Introduction

Though many encouraging results have been reported in the challenging tasks of classification and recognition especially with the development of the convolutional neural networks (CNN) (Krizhevsky, Sutskever, and Hinton 2012), recent researches (Long et al. 2016; Tsai et al. 2016) show that the problem of dataset bias (Torralba and Efros 2011) remained unsolved. Even with more generative features learnt by CNN, it can hardly ensure that the data distributions of the training dataset (source domain) and the test dataset (target domain) are exactly the same. Some techniques (Bruzzone and Marconcini 2010; Yao et al. 2015) have been developed to refine the source classifier for the target domain with target labeled data, but it is extremely expensive and time-consuming to annotate target data in practical applications (e.g., large scale camera networks). Instead of collecting target labels, unsupervised domain adaptation (Gopalan, Li, and Chellappa 2011) has been proposed to address the problem of distribution mismatch under the situation that target labels are unavailable. Existing methods can be categorized into two approaches as follows.

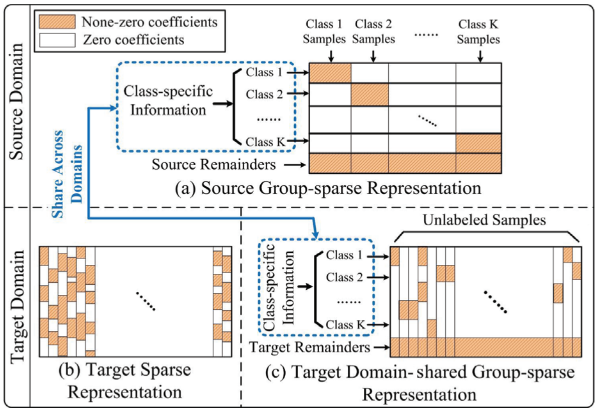

Figure 1: Illustration of Domain-shared Group-sparse Representation (a) Group-sparse representation for source labeled data; (b) Sparse representation for target unlabeled data; (c) Without target labels, we constrain the target coefficients share the same group sparsity as the source domain. As the computation of distribution does not rely on data order, conditional distributions across domains can be aligned by the proposed domain-shared group-sparse constraint.

The first approach assumes that conditional distributions between the source and target domains are equal, and therefore, joint distribution alignment can be solved by aligning marginal distributions across domains. Along this line, many unsupervised domain adaptation methods, e.g. (Sun, Feng, and Saenko 2016; Tsai et al. 2016; Tzeng et al. 2017), have been proposed and the results are encouraging when the assumption is valid. However, in many practical applications such as face and object recognition, conditional distributions between source and target domains may be different.

The second approach matches the joint distributions across domains without equal conditional distribution assumption. For this purpose, (Long et al. 2013b; Ming Harry Hsu et al. 2015) proposed to estimate target labels from unlabeled target data so that target conditional distribution can be estimated for joint distribution alignment. However, accurately estimating labels from unlabeled target data is very difficult due to the large data variations. Instead, (Long et al. 2014; Gong et al. 2016) re-weighted and/or transformed source data as target labeled data. Under the assumptions of location-scale transformation and conditional independence between different classes, they proved that the joint distributions across domains can be aligned. However, such assumptions may not be valid in practical applications.

In view of the limitations on existing methods, this paper proposes to derive an equivalent condition for conditional distribution alignment, instead of estimating target labeled data directly. In specific, we take advantage of the class-independent property in group-sparse representations (Sun et al. 2014) to extract class-specific information. By sharing the class-specific information across domains, the conditional distribution alignment can be modeled by constraining the source and target domains share the same group sparsity.

As shown in Figure 1 (a), source data are sparse in group according to their labels in the group-sparse representations. But this property cannot be guaranteed for target data without labels as shown in Figure 1 (b). With different non-zero elements in the source and target coefficient vectors of the same class, the conditional distributions are not the same. In contrast, constrained by domain-shared group sparsity, the source and target coefficient vectors have and only have non-zero elements in one class-specific group (as illustrated in Figure 1 (c)). Using the same group of elements for representation, the conditional distributions of data from each class are aligned in the group-sparse coefficients.

The contributions of this paper are three-folds,

- We derive a domain-shared group-sparse constraint for conditional distributions alignment in unsupervised domain adaptation.
- We propose a domain-shared group-sparse dictionary learning model to align joint distributions across domains.
- We develop a target-specific classifier that exploits both domain-shared and target-specific information.

2 Related Work

In unsupervised domain adaptation, many researches devoted to select or project source data as target labeled data to train a classifier for the target domain. Selection-based methods (Gong, Sha, and Grauman 2012; Gong, Grauman, and Sha 2013) selected a subset of source data, in which the distribution is similar to the target domain for adaptation. In contract, projection-based methods (Gong et al. 2012; Sun, Feng, and Saenko 2016; Tsai et al. 2016) projected the source data to the target domain or projected both source and target data into an intermediary space, so that the projected source dataset could be used as target-domain training data. Along this line, dictionary-based models (Ni, Qiu, and Chellappa 2013; Wu et al. 2016) are developed to find the projected space with sparse representation, and deep-learning based methods (Long et al. 2016; Ganin et al. 2016) are proposed to model the nonlinear projection.

Other methods aimed to estimate labels from target data for adaptation. JDA (Long et al. 2013b) predicted labels for target data with the source classifier to align data distributions across domains. (Ming Harry Hsu et al. 2015) further improved JDA by combing the structural information of target data to predict target labels. Rather than distribution alignment, (Xu et al. 2014) refined the source classifiers to the target domain using target data with estimated labels.

3 Proposed Method

In this section, we first introduce the Domain-shared Group-sparse Dictionary Learning (DsGsDL) model that aligns joint distributions across domains by minimizing the domain-shared group-sparse constraint and marginal distribution difference. Theoretical analysis and optimization algorithm are then provided in Section 3.2 and 3.3, respectively. After learning the domain-shared group-sparse dictionary, a classifier for the target domain is trained using the domain-shared coefficients together with target-specific information from target data. Details about the classifier learning is discussed in Section 3.4.

3.1 Domain-Shared Group-Sparse Dictionary Learning

Given a set of source data X^S with labels $\boldsymbol{y}^S$ and a set of unlabeled target data X^T, we aim to align joint distributions $P_S(X^S, \boldsymbol{y}^S)$ and $P_T(X^T, \boldsymbol{y}^T)$ for domain adaptation. Specifically, target labels $\boldsymbol{y}^T$ are unknown in unsupervised domain adaptation. For this purpose, the proposed DsGsDL model learns the group-sparse representations, in which both the conditional and marginal distributions are aligned across the source and target domains.

Conditional Distribution Alignment We align conditional distributions with the constraint of domain-shared group sparsity on both source and target domains. We first consider the formulation in the source domain. Denote a set of labeled source data from K classes as $X^S = [X_1^S, X_2^S, ..., X_K^S]$, where $X_k^S \in \mathbb{R}^{p \times n_k}$ is the sub-set of source data from class k, p is the feature dimensionality of each source sample and n_k is the number of source data from class k. The dictionary for source domain is represented as $D^S = [D_1^S, D_2^S, ..., D_K^S, D_r^S]$, where $D_k^S \in \mathbb{R}^{p \times q_k}$ is a sub-dictionary specific to class k and $D_r^S \in \mathbb{R}^{p \times q_r}$ is the dictionary for the remainder sparse coefficients from all classes in the source domain. q_k and q_r are the number of bases for D_k^S and D_r^S, respectively. Let $\boldsymbol{\alpha}^S \in \mathbb{R}^{q \times n}$ be the coefficients for source data. q is the total number of bases in D^S while n is the total number of source data. Correspond to D^S, source coefficients are divided as matrix of row vectors $\boldsymbol{\alpha}^S = [\boldsymbol{\alpha}_{1,:}^S; \boldsymbol{\alpha}_{2,:}^S; ...; \boldsymbol{\alpha}_{K,:}^S; \boldsymbol{\alpha}_{r,:}^S]$. On the other hand, the coefficient matrix can be written by column vectors as $\boldsymbol{\alpha}^S = [\boldsymbol{\alpha}_{:,1}^S, \boldsymbol{\alpha}_{:,2}^S, ..., \boldsymbol{\alpha}_{:,K}^S]$ according to the source labels $\boldsymbol{y}^S$. We attain source-domain group sparsity by minimizing the reconstruction error of each sub-dictionary and constraining that samples from different classes respond to different sub-dictionaries. Employing l_0 norm for group-sparse constraint, the source group-sparse dictionary is learned by

$$\begin{aligned} \min_{D^S, \boldsymbol{\alpha}^S} & \sum_{k=1}^{K} ||X_k^S - D_k^S \boldsymbol{\alpha}_{k,k}^S - D_r^S \boldsymbol{\alpha}_{r,k}^S||_F^2 \\ & + \eta \sum_{\boldsymbol{y}_i \neq \boldsymbol{y}_j}^{n} ||\boldsymbol{\alpha}_{c,(i)}^S \circ \boldsymbol{\alpha}_{c,(j)}^S||_0 + \lambda \sum_{i=1}^{n} |\boldsymbol{\alpha}_{(i)}^S| \end{aligned} \tag{1}$$

where $\circ$ denotes the Hadamard product, $\boldsymbol{\alpha}_c^S = [\boldsymbol{\alpha}_{1,:}^S; \boldsymbol{\alpha}_{2,:}^S; ...; \boldsymbol{\alpha}_{K,:}^S]$ is the class-specific coefficients, $\boldsymbol{\alpha}_{c,(i)}^S$ and $\boldsymbol{\alpha}_{c,(j)}^S$ are the i-th and j-th column of $\boldsymbol{\alpha}_c^S$, respectively. $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ represent the labels of the i-th and j-th source samples, respectively. $\boldsymbol{\alpha}_{(i)}^S$ is the coefficients for the i-th source sample. η and λ are regularization parameters for coefficient sparsity.

In the target domain, the domain-shared group sparsity is formulated as follows. Given a set of unlabeled target data $X^T \in \mathbb{R}^{p \times m}$, where p is feature dimensionality and m is the number of target data, let target dictionary has the same structure as source dictionary, denoted as $D^T = [D_1^T, D_2^T, ..., D_K^T, D_r^T]$. Correspondingly, denote target coefficients as $\boldsymbol{\alpha}^T = [\boldsymbol{\alpha}_c^T; \boldsymbol{\alpha}_r^T]$, where $\boldsymbol{\alpha}_c^T = [\boldsymbol{\alpha}_{1,:}^T; \boldsymbol{\alpha}_{2,:}^T; ...; \boldsymbol{\alpha}_{K,:}^T]$ is target class-specific coefficients. Since the classification task is shared across source and target domains, the sub-dictionary D_k^S specific to class k is close to the sub-dictionary D_k^T in target domain. Utilizing this criterion, the distance between class-specific components $D_c^S = [D_1^S, D_2^S, ..., D_K^S]$ and $D_c^T = [D_1^T, D_2^T, ..., D_K^T]$ in source and target domains are minimized for the same class representation across domains, i.e.,

$$\min_{D_c^S, D_c^T} ||D_c^S - D_c^T||_F^2 \tag{2}$$

On the other hand, the coefficients of the same class samples in both domains need to share the same group sparsity for conditional distribution alignment. For this purpose, the coefficients of target samples are constrained to exclude for two sub-dictionaries representing two different classes. With this criterion, each target sample automatically selects the optimal group of coefficients for representation so that the conditional distributions of source and target group-sparse coefficients are aligned. We employ l_0 norm to obtain group sparsity for target coefficients $\boldsymbol{\alpha}^T$, which minimize

$$\min_{\boldsymbol{\alpha}^T} \sum_{l_a \neq l_b}^{q_c} ||\boldsymbol{\alpha}_{(a),:}^T \circ \boldsymbol{\alpha}_{(b),:}^T||_0 \tag{3}$$

where $\boldsymbol{\alpha}_{(a),:}^T$ and $\boldsymbol{\alpha}_{(b),:}^T$ are the a-th and b-th rows of $\boldsymbol{\alpha}^T$, respectively. $q_c = \sum_{k=1}^K q_k$ is the total number of class-specific bases. l_a and l_b represent the class of sub-dictionaries that $\boldsymbol{\alpha}_{(a),:}^T$ and $\boldsymbol{\alpha}_{(b),:}^T$ respond, respectively.

Marginal Distribution Alignment Besides conditional distribution alignment, we further minimize the marginal distribution difference for joint distribution alignment. Maximum Mean Discrepancy (MMD) (Gretton et al. 2007) is employed to measure the difference between marginal distributions of the domain-shared group-sparse coefficients $\boldsymbol{\alpha}_c^S$ and $\boldsymbol{\alpha}_c^T$. Thus, we minimize

$$\min_{\boldsymbol{\alpha}_c} \boldsymbol{tr}(\boldsymbol{\alpha}_c M \boldsymbol{\alpha}_c') \tag{4}$$

where $\boldsymbol{\alpha}_c = [\boldsymbol{\alpha}_c^S, \boldsymbol{\alpha}_c^T]$ and M is the MMD matrix,

$$M_{ij} = \begin{cases} 1/n^2, & i,j \leq n \\ 1/m^2, & i,j > n \\ 1/m/n, & otherwise \end{cases} \tag{5}$$

Combining conditional distribution alignment objectives (Equation (1), (2), (3)) and marginal distribution alignment objective (Equation (4)) with the target reconstruction error, the optimization problem towards aligning the joint distributions is given by

$$\begin{aligned} \min_{\substack{D^S, D^T \\ \boldsymbol{\alpha}^S, \boldsymbol{\alpha}^T}} \quad & \sum_{k=1}^{K} ||X_k^S - D_k^S \boldsymbol{\alpha}_{k,k}^S - D_r^S \boldsymbol{\alpha}_{r,k}^S||_F^2 \\ & + ||X^T - D^T \boldsymbol{\alpha}^T||_F^2 + \eta \sum_{\boldsymbol{y}_i \neq \boldsymbol{y}_j}^{n} ||\boldsymbol{\alpha}_{c,(i)}^S \circ \boldsymbol{\alpha}_{c,(j)}^S||_0 \\ & + \delta \sum_{l_a \neq l_b}^{q_c} ||\boldsymbol{\alpha}_{(a),:}^T \circ \boldsymbol{\alpha}_{(b),:}^T||_0 + \beta ||D_c^S - D_c^T||_F^2 \\ & + \mu \boldsymbol{tr}(\boldsymbol{\alpha}_c M \boldsymbol{\alpha}_c') + \lambda \sum_{i=1}^{n+m} |\boldsymbol{\alpha}_{(i)}| \end{aligned} \tag{6}$$

where $\boldsymbol{\alpha}_{(i)}$ is the i-th column of $\boldsymbol{\alpha} = [\boldsymbol{\alpha}^S, \boldsymbol{\alpha}^T]$ and μ, β, η, λ and δ are parameters to balance the trade-off between the reconstruction terms and other constraints.

3.2 Theoretical Analysis

In our model, the joint distributions of the domain-shared group-sparse coefficients $P_S(\boldsymbol{\alpha}_c^S, \boldsymbol{y}^S)$ and $P_T(\boldsymbol{\alpha}_c^T, \boldsymbol{y}^T)$ are aligned by jointly aligning the marginal and conditional distributions across domains. Marginal distributions $P_S(\boldsymbol{\alpha}_c^S)$ and $P_T(\boldsymbol{\alpha}_c^T)$ are aligned by minimizing the MMD between the source and target domains (Equation (4)), as in existing methods (Long et al. 2013a). In the following, we mainly discuss how the conditional distributions can be aligned by the constraint of domain-shared group sparsity.

With labeled data in the source domain, group-sparse representations $\boldsymbol{\alpha}_c^S$ (Equation (1)) can be obtained by group-sparse dictionary learning (Sun et al. 2014). With the group-sparse representations in the source domain, the coefficients of source data from class k are more likely to be non-zeros for the k-th class-specific dictionary D_k^S, while others are almost zeros. With this property, we show that domain-shared group sparsity is a sufficient and necessary condition for conditional distribution alignment across the source and target domains as follows.

Sufficiency: Denote χ_k as a set of coefficient vectors in the k-th group (from class k). Source-domain group sparsity implies large-value probability $P_S(\boldsymbol{y}^S = k | \boldsymbol{\alpha}_c^S \subseteq \chi_k)$ and small-value probability $P_S(\boldsymbol{y}^S \neq k | \boldsymbol{\alpha}_c^S \subseteq \chi_k)$. With the same classification task, we assume that the class-specific information is shared across the source and target domains. According to the domain-shared property (Equation (2)), the k-th group target dictionary D_k^T is corresponding to the k-th class-specific dictionary D_k^S, which means the class concept is shared by source and target dictionaries across domains. On the other hand, Equation (3) limits non-zero coefficients of target data to be in only one group k, i.e., $\boldsymbol{\alpha}_{c,k}^T \subseteq \chi_k$. Thus, we have large-value probability $P_T(\boldsymbol{y}^T = k | \boldsymbol{\alpha}_c^T \subseteq \chi_k)$ and small-value probability $P_T(\boldsymbol{y}^T \neq k | \boldsymbol{\alpha}_c^T \subseteq \chi_k)$ in the target domain. Consequently,

$P_S(\boldsymbol{y}^S|\boldsymbol{\alpha}_c^S) \approx P_T(\boldsymbol{y}^T|\boldsymbol{\alpha}_c^T)$, which means the conditional distributions are aligned across the source and target domains. Hence, domain-shared group sparsity is a sufficient condition for conditional distribution alignment.

Necessity: In the source domain, according to the previous analysis, the group-sparse coefficients $\boldsymbol{\alpha}_c^S$ satisfies $\boldsymbol{\alpha}_c^S \subseteq (\cup_{k=1}^K \chi_k)$, which implies large-value probability $P_S(\boldsymbol{y}^S|\boldsymbol{\alpha}_c^S \subseteq (\cup_{k=1}^K \chi_k))$ and small-value probability $P_S(\boldsymbol{y}^S|\boldsymbol{\alpha}_c^S \not\subseteq (\cup_{k=1}^K \chi_k))$. If the conditional probabilities are equal to each other in both domains, we get large-value probability $P_T(\boldsymbol{y}^T|\boldsymbol{\alpha}_c^T \subseteq (\cup_{k=1}^K \chi_k))$ and small-value probability $P_T(\boldsymbol{y}^T|\boldsymbol{\alpha}_c^T \not\subseteq (\cup_{k=1}^K \chi_k))$. This means the group sparsity is shared across domains, so $\boldsymbol{\alpha}_c^T \subseteq (\cup_{k=1}^K \chi_k)$. Therefore, target coefficients $\boldsymbol{\alpha}_c^T$ have non-zero values in only one group as the source domain, i.e., $||\boldsymbol{\alpha}_{(a),:}^T \circ \boldsymbol{\alpha}_{(b),:}^T||_0 = 0$ for $l_a \neq l_b$ (Equation (3)). On the other hand, conditional distributions are equal to each other for each class, i.e., $P_S(\boldsymbol{y}^S = k|\boldsymbol{\alpha}_c^S \subseteq \chi_k) = P_T(\boldsymbol{y}^T = k|\boldsymbol{\alpha}_c^T \subseteq \chi_k)$. This constrains that class information encoded in the dictionaries is shared across domains, so we formulate this property as $D_c^S = D_c^T$ (Equation (2)). Therefore, domain-shared group sparsity is a necessary condition for equal conditional distribution across the source and target domains.

3.3 Optimization

To solve the optimization problem in Equation (6) more efficiently, we introduce a set of selection matrices $\{Q^1, Q^2, ...Q^K, Q^r\}$ to extract sub-dictionaries from source and target dictionaries. Let $Q^k \in \mathbb{R}^{q\times q} = diag(0,1)$ be a diagonal matrix, where the diagonal value $Q_{ii}^k = 1$ if the i-th basis of D^S and D^T belongs to sub-dictionary D_k^S and D_k^T, respectively. Similarly, $Q^r \in \mathbb{R}^{q\times q} = diag(0,1)$ has none-zero values in the position corresponding to D_r^S and D_r^T. Applying selection matrices, we have $||X_k^S - D_k^S\boldsymbol{\alpha}_{k,k}^S - D_r^S\boldsymbol{\alpha}_{r,k}^S||_F^2 = ||X_k^S - D^S Q^k \boldsymbol{\alpha}_{:,k}^S - D^S Q^r \boldsymbol{\alpha}_{:,k}^S||_F^2$. Let $Q^c = \sum_{k=1}^K Q^k$ be the selection matrix for class-specific dictionaries, we get $D_c^S = D^S Q^c$, $D_c^T = D^T Q^c$, $\boldsymbol{\alpha}_c = Q^c\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_{c,(i)} = Q^c\boldsymbol{\alpha}_{(i)}$. Equation (6) is then rewritten as,

$$\begin{aligned}\min_{\substack{D^S, D^T\\ \boldsymbol{\alpha}^S, \boldsymbol{\alpha}^T}} \quad & \sum_{k=1}^{K} ||X_k^S - D^S Q^k \boldsymbol{\alpha}_{:,k}^S - D^S Q^r \boldsymbol{\alpha}_{:,k}^S||_F^2 \\ & + ||X^T - D^T\boldsymbol{\alpha}^T||_F^2 + \eta \sum_{\boldsymbol{y}_i \neq \boldsymbol{y}_j}^{n} ||Q^c\boldsymbol{\alpha}_{(i)}^S \circ Q^c\boldsymbol{\alpha}_{(j)}^S||_0 \\ & + \delta \sum_{l_a \neq l_b}^{q_c} ||\boldsymbol{\alpha}_{(a),:}^T \circ \boldsymbol{\alpha}_{(b),:}^T||_0 + \mu \boldsymbol{tr}(Q^c\boldsymbol{\alpha} M (Q^c\boldsymbol{\alpha})') \\ & + \beta ||D^S Q^c - D^T Q^c||_F^2 + \lambda \sum_{i=1}^{n+m} |\boldsymbol{\alpha}_{(i)}| \end{aligned} \tag{7}$$

To solve the optimization problem (7), dictionaries D^S and D^T, coefficients $\boldsymbol{\alpha}^S$ and $\boldsymbol{\alpha}^T$ are iteratively optimized as follow.

Learning Dictionary Fix source and target coefficients $\boldsymbol{\alpha}^S$, $\boldsymbol{\alpha}^T$, source dictionary D^S is learned by solving the l_2 norm optimization function,

$$\begin{aligned}\min_{D^S} \quad & \sum_{k=1}^{K} ||X_k^S - D^S(Q^k\boldsymbol{\alpha}_{:,k}^S - Q^r\boldsymbol{\alpha}_{:,k}^S)||_F^2 \\ & + \beta||D^S Q^c - D^T Q^c||_F^2 \end{aligned} \tag{8}$$

The solution to (8) can be obtained by setting the first derivative to zero. Similarly, fix $\boldsymbol{\alpha}^S$, $\boldsymbol{\alpha}^T$ and D^S, target dictionary D^T is learned by

$$\min_{D^T} \quad ||X^T - D^T\boldsymbol{\alpha}^T||_F^2 + \beta||D^S Q^c - D^T Q^c||_F^2 \tag{9}$$

Learning Coefficients The strategy of coordinate descent (Aharon, Elad, and Bruckstein 2006) is utilized to update each column $\boldsymbol{\alpha}_{(i)}$ of coefficients $\boldsymbol{\alpha} = [\boldsymbol{\alpha}^S, \boldsymbol{\alpha}^T]$ when D^S, D^T and other columns of $\boldsymbol{\alpha}$ are fixed.

For the source coefficient $\boldsymbol{\alpha}_{(i)}$, where $i \leq n$, the optimization function is,

$$\min_{\boldsymbol{\alpha}_{(i)}} g^S(\boldsymbol{\alpha}_{(i)}) = h^S(\boldsymbol{\alpha}_{(i)}) + l^S(\boldsymbol{\alpha}_{(i)}) + \lambda|\boldsymbol{\alpha}_{(i)}| \tag{10}$$

where

$$\begin{aligned} h^S(\boldsymbol{\alpha}_{(i)}) =& ||\boldsymbol{x}_i - (D^S Q^{k_i} - D^S Q^r)\boldsymbol{\alpha}_{(i)}^S||_F^2 \\ & + \mu M_{ii}(Q^c\boldsymbol{\alpha}_{(i)})'(Q^c\boldsymbol{\alpha}_{(i)}) \\ & + 2\mu \sum_{j=1, j\neq i}^{n+m} M_{ij}(Q^c\boldsymbol{\alpha}_{(i)})'(Q^c\boldsymbol{\alpha}_{(j)}) \end{aligned} \tag{11}$$

$\boldsymbol{x}_i$ is the i-th column of $X = [X^S, X^T]$, k_i is the class index of $\boldsymbol{x}_i$ and

$$l^S(\boldsymbol{\alpha}_{(i)}) = \eta \sum_{\substack{j=1\\ \boldsymbol{y}_i \neq \boldsymbol{y}_j}}^{n} ||Q^c\boldsymbol{\alpha}_{(i)}^S \circ Q^c\boldsymbol{\alpha}_{(j)}^S||_0 \tag{12}$$

Minimizing Equation (10) is NP hard as it includes a l_0 norm term and a l_1 norm term. To solve the l_0 norm, an iterative re-weighting strategy (Chartrand and Yin 2008) is employed to approximate $l^S(\boldsymbol{\alpha}_{(i)})$. Follow the support discrimination method (Liu et al. 2016), $l^S(\boldsymbol{\alpha}_{(i)})$ is approximated by l_2 norm as,

$$l^S(\boldsymbol{\alpha}_{(i)}) = \eta||\varOmega_i Q^c \boldsymbol{\alpha}_{(i)}^S||_F^2 \tag{13}$$

where $\varOmega_i$ is a re-weight matrix calculated by coefficients with different class label from $\boldsymbol{\alpha}_{(i)}^S$.

$$\varOmega_i = diag(\sqrt{\sum_{j, \boldsymbol{y}_i \neq \boldsymbol{y}_j} (\sqrt{\boldsymbol{\omega}_{ij}} \circ (Q^c\boldsymbol{\alpha}_{(j)}^S))^{\odot 2}}) \tag{14}$$

with $\boldsymbol{\omega}_{ij} = 1/(Q^c\boldsymbol{\alpha}_{(i)}^S \circ Q^c\boldsymbol{\alpha}_{(j)}^S)^{\odot 2}$, where $\odot 2$ represents the operation of element by element square for a vector.

With Equation (13), Equation (10) is approximated as a QP problem with a sparsity constraint. Follow the optimization process proposed in TSC (Long et al. 2013a), a sparse coding algorithm (Lee et al. 2007) is applied. By searching the sign for each element of $\boldsymbol{\alpha}_{(i)}$, Equation (10) is reduced into a quadratic optimization problem, which can be solved by fixing the first derivative as zero.

For the target coefficient $\boldsymbol{\alpha}_i$, where $i > n$, the optimization function is

$$\min_{\boldsymbol{\alpha}_{(i)}} g^T(\boldsymbol{\alpha}_{(i)}) = h^T(\boldsymbol{\alpha}_{(i)}) + l^T(\boldsymbol{\alpha}_{(i)}) + \lambda|\boldsymbol{\alpha}_{(i)}| \tag{15}$$

where

$$\begin{aligned} h^T(\boldsymbol{\alpha}_{(i)}) =& ||\boldsymbol{x}_{(i)} - D^T\boldsymbol{\alpha}_{(i)}||_F^2 + \mu M_{ii}(Q^c\boldsymbol{\alpha}_{(i)})'(Q^c\boldsymbol{\alpha}_{(i)}) \\ &+ 2\mu \sum_{j=1, j\neq i}^{n+m} M_{ij}(Q^c\boldsymbol{\alpha}_{(i)})'(Q^c\boldsymbol{\alpha}_{(j)}) \end{aligned} \tag{16}$$

and

$$l^T(\boldsymbol{\alpha}_{(i)}) = \delta \sum_{l_a \neq l_b}^{q_c} ||\boldsymbol{\alpha}^T_{(a),(i)} \circ \boldsymbol{\alpha}^T_{(b),(i)}||_0 \tag{17}$$

We adapt the feature-sign search algorithm (Lee et al. 2007) to solve the NP hard problem of Equation (15) with a non-derivable term (Equation (17)). For the a-th element of $\boldsymbol{\alpha}^T_{(i)}$, we have

$$||\boldsymbol{\alpha}^T_{(a),(i)} \circ \boldsymbol{\alpha}^T_{(b)(i)}||_0 = \begin{cases} ||\boldsymbol{\alpha}^T_{(b)(i)}||_0, \; \boldsymbol{\alpha}^T_{(a),(i)} \neq 0 \\ \qquad 0, \; otherwise \end{cases} \tag{18}$$

If $\boldsymbol{\alpha}^T_{(a),(i)} \neq 0$, the subdifferentiable of $|\boldsymbol{\alpha}_{(i)}|$ for the a-th element is $\nabla_{(a)}|\boldsymbol{\alpha}_{(i)}| = sign(\boldsymbol{\alpha}_{(a),(i)})$. Thus, the subdifferentiable for $g^T(\boldsymbol{\alpha}_{(i)})$ is

$$\begin{aligned} \nabla_{(a)} g^T(\boldsymbol{\alpha}_{(i)}) =& \nabla_{(a)} h^T(\boldsymbol{\alpha}_{(i)}) + \lambda sign(\boldsymbol{\alpha}_{(a),(i)}) \\ &+ \delta \sum_{\substack{b=1 \\ l_a \neq l_b}}^{q_c} ||\boldsymbol{\alpha}^T_{(b),(i)}||_0 \end{aligned} \tag{19}$$

Otherwise, $\boldsymbol{\alpha}^T_{(a),(i)} = 0$, $\nabla_{(a)}|\boldsymbol{\alpha}_{(i)}|$ and $\nabla_{(a)}||\boldsymbol{\alpha}^T_{(a),(i)} \circ \boldsymbol{\alpha}^T_{(b)(i)}||_0$ are non-derivable so that we set the value for them as 1 or -1. The subdifferentiable for $g^T(\boldsymbol{\alpha}_{(i)})$ is

$$\nabla_{(a)} g^T(\boldsymbol{\alpha}_{(i)}) = \nabla_{(a)} h^T(\boldsymbol{\alpha}_{(i)}) \pm (\lambda + \delta) \tag{20}$$

To optimize Equation (15), the gradient $\nabla_{(a)} g^T(\boldsymbol{\alpha}_{(i)})$ should be equal to 0 if $\boldsymbol{\alpha}^T_{(a),(i)} \neq 0$. Otherwise, $|\nabla_{(a)} h^T(\boldsymbol{\alpha}_{(i)})| \leq \lambda + \delta$ and $\nabla_{(a)} g^T(\boldsymbol{\alpha}_{(i)})$ can not be 0 for any value of $\boldsymbol{\alpha}^T_{(a),(i)}$. Here, both λ and δ are positive numbers. Therefore, we search the elements of $\boldsymbol{\alpha}^T_{(i)}$ that satisfies $|\nabla_{(a)} h^T(\boldsymbol{\alpha}_{(i)})| > \lambda + \delta$ for optimization. For the a-th element of $\boldsymbol{\alpha}^T_{(i)}$, if $|\nabla_{(a)} h^T(\boldsymbol{\alpha}_{(i)})| > \lambda + \delta$, we get $\boldsymbol{\alpha}^T_{(a),(i)} \neq 0$ and assign sign for $\boldsymbol{\alpha}^T_{(a),(i)}$. The value of $\boldsymbol{\alpha}^T_{(a),(i)}$ is then optimized by setting Equation (19) equal to 0.

3.4 Learning Target Classifier by Incorporating Target-Domain-Specific Information

After getting the domain-shared group-sparse representations, a straight-forward method to classify target samples is to train a classifier using the source domain-shared coefficients $\boldsymbol{\alpha}^S_c$ with their labels $\boldsymbol{y}^S$, and then apply it to target coefficients $\boldsymbol{\alpha}^T_c$. While $\boldsymbol{\alpha}^S_c$ and $\boldsymbol{\alpha}^T_c$ only contain the discriminative information that is shared in both source and target domains, it is possible that some target discriminative information is not contained in the source domain.

To further incorporate target-domain-specific information, we propose to learn a target classifier by using both the domain-shared group-sparse representations and the target feature vectors X^T. On the one hand, source label information $\boldsymbol{y}^S$ is propagated to target-domain data through the domain-shared group-sparse representations, which is formulated similar to the objective in semi-supervised learning (e.g., (Wang and Zhang 2008)). On the other hand, the original feature space X^T is more informative than the domain-shared group-sparse coefficients, so the regularization graph within the target data is constructed by X^T.

Based on the analysis above, we construct two graphs for cross-domain label propagation. The first one $G^c =< V^c, E^c >$ is constructed by class-specific coefficients $\boldsymbol{\alpha}_c$ in both source and target domains, where the nodes $V^c = \boldsymbol{\alpha}_c$ and the edge $\boldsymbol{\varepsilon}^c_{ij} \in E^c$ connects $\boldsymbol{\alpha}_{c,(i)}$ and $\boldsymbol{\alpha}_{c,(j)}$. The second one $G^T =< V^T, E^T >$ is constructed by target data X^T only, where the nodes $V^T = X^T$ and the edge $\boldsymbol{\varepsilon}^T_{ij} \in E^T$ connects $\boldsymbol{x}^T_i$ and $\boldsymbol{x}^T_j$.

The weight of each edge between a pair of nodes is computed by the neighbourhood information of each node (Wang and Zhang 2008). With weight matrices W^c and W^T, the classifier in the target domain can be learned by label propagation on these two graphs. For simplicity, linear classifier $\boldsymbol{\theta}$ is employed, i.e.,

$$\begin{aligned} \min_{\boldsymbol{\theta}} \; & \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} ||\boldsymbol{y}^S_i - \boldsymbol{\theta}\boldsymbol{x}^{\boldsymbol{T}}_{\boldsymbol{j}}||_2^2 W^c_{j'i} \\ &+ \frac{1}{m^2}\sum_{i,j=1}^{m} ||\boldsymbol{\theta}\boldsymbol{x}^T_i - \boldsymbol{\theta}\boldsymbol{x}^T_j||_2^2 W^T_{ij} + \gamma||\boldsymbol{\theta}||_2^2 \end{aligned} \tag{21}$$

where $j' = j + n$ is the index of $\boldsymbol{\alpha}^T_{c,(j)}$ in $\boldsymbol{\alpha}_c$, and γ is a regularization parameter. The label of target data $\boldsymbol{x}^T_j$ depends on its neighbourhood in $\boldsymbol{\alpha}_c$ and X^T.

Equation (21) can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{\theta}} & \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\theta}\boldsymbol{x}^T_j(\boldsymbol{x}^T_j)'\boldsymbol{\theta}' - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(\boldsymbol{y}^S_i(\boldsymbol{x}^T_j)'\boldsymbol{\theta}')W^c_{j'i} \\ &+ \frac{1}{m^2}\boldsymbol{\theta}X^T L(X^T)'\boldsymbol{\theta}' + \gamma\boldsymbol{\theta}\boldsymbol{\theta}' \end{aligned} \tag{22}$$

where $L = diag(\sum_{j=1}^{m} W^T_{ij}) - W^T$ is the Laplacian matrix.

By calculating the derivation for Equation (22), the solution to the above optimization problem is given by $\boldsymbol{\theta} = (\frac{1}{nm}\sum_i\sum_j(\boldsymbol{y}^S_i(\boldsymbol{x}^T_j)')W^c_{j'i})(\frac{1}{m}\sum_j \boldsymbol{x}^T_j(\boldsymbol{x}^T_j)' + \frac{1}{m^2}X^T L(X^T)' + \gamma I)^{-1}$.

4 Experimental Results

In this section, we evaluate our method on two recognition tasks: 1) face recognition across blur and illumination; 2) object recognition across different datasets.

Table 1: Comparison of face recognition accuracy (%) on CMU-PIE dataset *(Source: images under 11 illumination conditions; Target: images under other 10 illumination conditions with a blur kernel)*

Method	σ =3	σ =4	L =9	L =11
SVM$_{Scr}$	73.82	70.00	72.94	67.35
GFK	77.65	74.71	80.88	73.53
SA	76.47	75.29	78.24	75.29
SIDL	71.47	68.53	73.53	64.41
TKL	77.94	77.35	78.24	76.76
CORAL	76.18	72.94	78.53	67.06
JDA	74.12	74.71	78.53	62.65
TJM	59.71	56.76	60.29	42.06
CTC	68.82	65.28	70.00	62.65
DsGsDL$_{NT}$ (Ours)	**86.47**	**86.76**	**89.12**	**77.74**
DsGsDL (Ours)	**88.82**	**87.94**	**90.59**	**80.29**

Table 2: Comparison of face recognition accuracy (%) on CMU-PIE dataset *(Source: images under 10 illumination conditions with a blur kernel; Target: images under other 11 illumination conditions)*

Method	σ =3	σ =4	L =9	L =11
SVM$_{Scr}$	70.88	66.48	74.12	67.06
GFK	85.29	83.24	85.29	81.47
SA	75.00	71.76	76.47	77.35
SIDL	83.53	80.59	87.06	82.06
TKL	80.88	80.59	81.76	77.35
CORAL	83.24	81.18	84.12	81.76
JDA	87.35	83.53	86.18	76.76
TJM	50.59	51.47	56.47	40.88
CTC	74.74	73.82	77.06	72.06
DsGsDL$_{NT}$ (Ours)	**98.53**	**98.35**	**97.94**	**97.35**
DsGsDL (Ours)	**98.82**	**99.41**	**98.24**	**98.82**

4.1 Experimental Settings

Implementation Details For learning the domain-shared group-sparse dictionary, the sizes of each sub-dictionary are 5 and 10 for face and object recognition, respectively. We set the number of remaining bases as 2 for experiments. The sparsity coefficient λ is set as 0.1, which follows the experimental settings in TSC (Long et al. 2013a). Hyper-parameter experiments are done to analyze the parameter sensitivity of parameters η, δ, μ and β in the optimization function (Equation (6)). To save the iteration times, the source-domain sparse dictionary (Equation (1)) are learnt to initialize the source and target dictionaries in the DsGsDL model (Equation (6)). Experiments show that our algorithm converges within 20-25 iterations during optimization.

For learning target classifier, the parameter γ is set as 1 to balance each term in Equation (21). The settings for learning the relationship graphs G^c and G^T follow the settings in (Wang and Zhang 2008). Target domain accuracy is used as performance measure like many existing domain adaptation methods (Gong et al. 2012; Long et al. 2013b).

Datasets CMU-PIE dataset (Sim, Baker, and Bsat 2002) is used for experiments of face recognition across blur and illumination variations. Following the protocol in (Ni, Qiu, and Chellappa 2013), frontal face images of 34 subjects under 21 illumination conditions are selected. The pixel values of each image are normalized and formed as a feature vector with 1920 dimensions for experiments. Two experiments are performed. In the first experiment, the source domain includes images under 11 different illumination conditions and target domain consists of the images under the other 10 illumination conditions with a blur kernel. Gaussian blur with standard deviations of 3 and 4 and motion blur with length of 9 and 11 are performed, respectively. The second experiment is implemented by reversing the source and target domains in the first experiment.

The experiments of object recognition are conducted on *Office+Caltech* dataset (Saenko et al. 2010), which contains four datasets of object images captured under different conditions: 1) *Amazon* includes 958 images downloaded from websites; 2) *Webcam* is make up by 295 low-resolution images taken by web camera; 3) *DSLR* contains 157 high-resolution images from digital SLR camera; 4) *Caltech-10* is a subset of a object recognition dataset *Caltech-256* with 1123 images. There are 10 categories of objects in each dataset, including bags, bikes, chairs, laptops, etc. For short, characters $\boldsymbol{A}$, $\boldsymbol{W}$, $\boldsymbol{D}$ and $\boldsymbol{C}$ are used to represent *Amazon*, *Webcam*, *DSLR*, and *Caltech-10*, respectively. We perform experiments on DeCAF$_6$ features (Donahue et al. 2014) with 4096 dimensions. Selecting one dataset as the source domain and any other dataset as the target domain, 4*3 pairs of source and target datasets are obtained for experiments.

Compared methods We compare the proposed method with eight state-of-the-art unsupervised domain adaptation methods, namely GFK (Gong et al. 2012), SA (Fernando et al. 2013), SIDL (Ni, Qiu, and Chellappa 2013), JDA (Long et al. 2013b), TJM (Long et al. 2014), TKL (Long et al. 2015b), CTC (Gong et al. 2016) and CORAL (Sun, Feng, and Saenko 2016). Among these, GFK, SA, SIDL, TKL and CORAL aligned marginal distributions under the assumption of equal conditional distribution across domains while JDA, TJM and CTC aligned both marginal and conditional distributions. In addition, results of two deep-learning based unsupervised domain adaptation methods (DDC (Tzeng et al. 2014) and DAN (Long et al. 2015a)) are presented for comparison in the experiments of cross-domain object recognition. For fair comparison, experimental settings of each compared method follow that reported in their paper. The results of the source-domain SVM classifier (SVM$_{Scr}$) are regarded as the baseline. To evaluate the performance of DsGsDL model, we also perform experiments on the DsGsDL model without using target-specific information, called DsGsDL$_{NT}$, in which SVM is used as a classifier.

Table 3: Comparison of classification performance (%) on Office+Caltech dataset with DeCAF$_6$ features *(Color **red**, **blue** and **green** represent the **best**, **second best** and **third best** results for each pair of datasets, respectively)*

Datasets	SVM$_{Scr}$	GFK	SA	SIDL	TKL	CORAL	JDA	TJM	CTC	DsGsDL$_{NT}$ (Ours)	DsGsDL (Ours)
$\boldsymbol{D \rightarrow W}$	**99.32**	97.29	98.22	**99.66**	98.31	98.98	**99.32**	97.97	97.29	**99.32**	**100**
$\boldsymbol{D \rightarrow A}$	83.19	81.63	81.06	85.18	77.04	86.01	**90.50**	86.95	86.12	**89.46**	**91.65**
$\boldsymbol{D \rightarrow C}$	76.67	75.42	76.06	80.77	74.09	77.47	**84.06**	76.76	74.09	**82.55**	**85.40**
$\boldsymbol{W \rightarrow D}$	99.36	96.82	98.69	**100**	**100**	99.36	99.36	**100**	99.36	**100**	**100**
$\boldsymbol{W \rightarrow A}$	81.21	87.89	77.21	82.25	76.62	82.36	**90.50**	85.91	78.39	**88.10**	**91.54**
$\boldsymbol{W \rightarrow C}$	74.80	77.29	72.62	75.33	71.15	75.78	**83.79**	75.42	74.09	**81.48**	**84.51**
$\boldsymbol{A \rightarrow D}$	82.80	84.08	81.31	71.97	78.98	80.89	77.71	**85.99**	82.80	**89.81**	**93.63**
$\boldsymbol{A \rightarrow W}$	77.97	**82.03**	74.64	65.76	76.61	74.57	80.34	80.00	73.22	**83.73**	**87.46**
$\boldsymbol{A \rightarrow C}$	82.99	**83.88**	81.90	81.21	80.05	83.43	80.94	78.01	78.98	**85.57**	**87.44**
$\boldsymbol{C \rightarrow D}$	87.26	84.71	83.50	76.43	87.26	87.90	80.89	**93.63**	79.62	**95.54**	**97.45**
$\boldsymbol{C \rightarrow W}$	80.34	**86.10**	76.15	71.53	71.19	80.00	82.37	81.69	72.20	**94.92**	**96.27**
$\boldsymbol{C \rightarrow A}$	90.81	91.65	89.81	85.70	91.02	91.65	86.74	**92.38**	87.47	**93.01**	**93.22**
Average	84.73	85.73	82.60	81.32	82.79	84.87	85.45	**86.23**	81.97	**90.29**	**92.38**

4.2 Face Recognition Across Blur and Illumination Variations

Experimental results of the two experiments on face recognition across blur and illumination are shown in Table 1 and Table 2, respectively. From these results, we can see that even without target-specific information, the proposed method (DsGsDL$_{NT}$) outperforms eight state-of-the-art unsupervised domain adaption methods in all variation settings. DsGsDL$_{NT}$ obtains $\mathbf{85.02}\%$ and $\mathbf{98.04}\%$ of average accuracy from the four different variation settings in the target domain for the first and second experiments, respectively. Incorporating with target-specific information, the proposed method (DsGsDL) achieves even higher accuracy, getting around $\mathbf{86.91}\%$ and $\mathbf{98.82}\%$ of accuracy at average in the first and second experiments, respectively. It can be observed from Table 1 and Table 2 that the proposed method (DsGsDL) gains a significant improvement (over $\mathbf{14.19}\%$ and $\mathbf{38.24}\%$ in Table 1 and Table 2, respectively) compared to the baseline (SVM$_{Scr}$) for recognizing faces in the target domain.

Moreover, it is interesting to see that the existing joint distribution alignment methods (JDA, TJM and CTC) do not always outperform the domain adaptation methods with equal conditional distribution assumption (GFK, SA, SIDL, TKL and CORAL). Especially, the performance of TJM which tries to match the conditional distributions by re-weighting (or selecting) source domain data is even worse than the source-domain classifier SVM$_{Scr}$. These results reflect that the selected source data can not correctly represent the target conditional distribution in cross-domain face recognition.

4.3 Object Recognition Across Datasets

Results of object recognition across datasets are recorded in Table 3. As shown in this table, without using target-specific information, DsGsDL$_{NT}$ outperforms eight state-of-the-art unsupervised domain adaption methods on 7 (out of 12) pairs of datasets. Higher accuracy (rank first in all datasets) is achieved by DsGsDL with target-specific information, which convinces that recognition performance can be further improved by incorporating the target-specific information. The average accuracy of our method (DsGsDL) on 12 pairs of datasets is $\mathbf{92.38}\%$, which is ($\mathbf{7.65}\%$) higher than the baseline (SVM$_{Scr}$).

Table 4: Average accuracy (%) of object recognition on Office+Caltech dataset compared with deep-learning based unsupervised domain adaptation methods

Method	DDC	DAN	DsGsDL (Ours)
Average accuracy	88.16	91.18	**92.38**

Besides, we find that the average accuracy of SVM$_{Scr}$ is better than some domain adaptation methods with equal conditional distribution assumption. This may indicate that the equal conditional distribution assumption is not valid in some cases of cross-domain object recognition. By matching conditional distributions, the joint distribution alignment methods (JDA and TJM) outperform SVM$_{Scr}$ and most of other domain adaptation methods except ours. Since our method avoids the problem of inaccuracy label estimation or source domain data selection, it achieves the best results.

Furthermore, we also compare the results of the proposed method (DsGsDL) to the deep-learning based unsupervised domain adaptation methods (DDC and DAN). As shown in Table 4, our method (DsGsDL) also achieves better result than DDC and DAN at average.

5 Conclusion

This paper proposes and develops a new Domain-shared Group-sparse Dictionary Learning method to align the joint distributions for unsupervised domain adaptation. With the domain-shared group-sparse coefficients, the target classifier is determined by further incorporating the target specific information. Experiments show that the domain-shared group-sparse coefficients achieve promising results in the tasks of cross-domain face and object recognition. The performance is further improved by incorporating target-specific information to train a classifier for the target domain.

Acknowledgements

This work was supported by the Science Faculty Research Grant of Hong Kong Baptist University, Hong Kong Research Grants Council General Research Fund: *RGC/HKBU12202514* and SYSU Research Funding: *67000-18821116*.

References

Aharon, M.; Elad, M.; and Bruckstein, A. 2006. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *TSP*.

Bruzzone, L., and Marconcini, M. 2010. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *TPAMI*.

Chartrand, R., and Yin, W. 2008. Iteratively reweighted algorithms for compressive sensing. In *ICASSP*.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*.

Fernando, B.; Habrard, A.; Sebban, M.; and Tuytelaars, T. 2013. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR*.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*.

Gong, M.; Zhang, K.; Liu, T.; Tao, D.; Glymour, C.; and Schölkopf, B. 2016. Domain adaptation with conditional transferable components. In *ICML*.

Gong, B.; Grauman, K.; and Sha, F. 2013. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*.

Gong, B.; Sha, F.; and Grauman, K. 2012. Overcoming dataset bias: An unsupervised domain adaptation approach. In *NIPS Workshop*.

Gopalan, R.; Li, R.; and Chellappa, R. 2011. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2007. A kernel method for the two-sample-problem. In *NIPS*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Lee, H.; Battle, A.; Raina, R.; and Ng, A. Y. 2007. Efficient sparse coding algorithms. In *NIPS*.

Liu, Y.; Chen, W.; Chen, Q.; and Wassell, I. 2016. Support discrimination dictionary learning for image classification. In *ECCV*.

Long, M.; Ding, G.; Wang, J.; Sun, J.; Guo, Y.; and Yu, P. S. 2013a. Transfer sparse coding for robust image representation. In *CVPR*.

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2013b. Transfer feature learning with joint distribution adaptation. In *ICCV*.

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2014. Transfer joint matching for unsupervised domain adaptation. In *CVPR*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015a. Learning transferable features with deep adaptation networks. In *ICML*.

Long, M.; Wang, J.; Sun, J.; and Philip, S. Y. 2015b. Domain invariant transfer kernel learning. *TKDE*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *NIPS*.

Ming Harry Hsu, T.; Yu Chen, W.; Hou, C.-A.; Hubert Tsai, Y.-H.; Yeh, Y.-R.; and Frank Wang, Y.-C. 2015. Unsupervised domain adaptation with imbalanced cross-domain data. In *ICCV*.

Ni, J.; Qiu, Q.; and Chellappa, R. 2013. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *CVPR*.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *ECCV*.

Sim, T.; Baker, S.; and Bsat, M. 2002. The cmu pose, illumination, and expression (pie) database. In *ICAFGR*.

Sun, Y.; Liu, Q.; Tang, J.; and Tao, D. 2014. Learning discriminative dictionary for group sparse representation. *TIP*.

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

Torralba, A., and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*.

Tsai, Y.-H. H.; Hou, C.-A.; Chen, W.-Y.; Yeh, Y.-R.; and Wang, Y.-C. F. 2016. Domain-constraint transfer coding for imbalanced unsupervised domain adaptation. In *AAAI*.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. *CVPR*.

Wang, F., and Zhang, C. 2008. Label propagation through linear neighborhoods. *TKDE*.

Wu, S.; Jing, X.-Y.; Yue, D.; Zhang, J.; Yang, K. J.; and Yang, J. 2016. Unsupervised visual domain adaptation via dictionary evolution. In *ICME*.

Xu, J.; Ramos, S.; Vázquez, D.; and López, A. M. 2014. Domain adaptation of deformable part-based models. *TPAMI*.

Yao, T.; Pan, Y.; Ngo, C.-W.; Li, H.; and Mei, T. 2015. Semi-supervised domain adaptation with subspace learning for visual recognition. In *CVPR*.