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

Information-Theoretic Domain Adaptation under Severe Noise Conditions

Wei Wang,[1] Hao Wang,[2] Zhi-Yong Ran,[3] Ran He[4]*
[1]Institute of Software, Chinese Academy of Sciences, Beijing 100190, China.
[2]360 Search Lab, Qihoo 360, Beijing 100190, China.
[3]Chongqing University of Posts and Telecommunications, Chongqing 400065, China.
[4]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
wangwei2014@iscas.ac.cn, cashenry@126.com, ranzy@cqupt.edu.cn, rhe@nlpr.ia.ac.cn

Abstract

Cross-domain data reconstruction methods derive a shared transformation across source and target domains. These methods usually make a specific assumption on noise, which exhibits limited ability when the target data are contaminated by different kinds of complex noise in practice. To enhance the robustness of domain adaptation under severe noise conditions, this paper proposes a novel reconstruction based algorithm in an information-theoretic setting. Specifically, benefiting from the theoretical property of correntropy, the proposed algorithm is distinguished with: detecting the contaminated target samples without making any specific assumption on noise; greatly suppressing the negative influence of noise on cross-domain transformation. Moreover, a relative entropy based regularization of the transformation is incorporated to avoid trivial solutions with the reaped theoretic advantages, i.e., non-negativity and scale-invariance. For optimization, a half-quadratic technique is developed to minimize the non-convex information-theoretic objectives with explicitly guaranteed convergence. Experiments on two real-world domain adaptation tasks demonstrate the superiority of our method.

Introduction

The task of domain adaptation refers to transferring knowledge from a well-learned source domain with sufficient labeled data to a target domain, where the two domains follow different but related distributions (Pan and Yang 2010). It plays a substantial role to the success of supervised learning machines when there are no or insufficient labeled training data in the target domain due to the expensive hand-labeling. As one of the most important families in domain adaptation, feature extraction (Pan et al. 2011; Hoffman et al. 2014) has attracted great attention due to their witnessed promising applications in multi-orientation face recognition, heterogeneous object classification, etc.

Generally, the key challenge in the feature extraction family of domain adaptation is to seek a shared feature space across the domains where the discrepancy of two distributions is minimized while the desired data properties are preserved. For instance, a representative subcategory employs a measurement to estimate and reduce the distance between the distributions (Si, Tao, and Geng 2010; Long et al. 2013). Another line of work bridges the distribution gap by building a deep structure (Zhou et al. 2014). Within the feature extraction family in domain adaptation, considerable efforts have also been devoted to cross-domain data reconstruction (Shao, Kit, and Fu 2014; Ding, Shao, and Fu 2015; Xu et al. 2016; Zhang, Zuo, and Zhang 2016). In this reconstruction based subcategory, the methods are more concerned on uncovering the intrinsic structures within two domains and modeling the noise in target domain during adaptation, showing robust potentials. Specifically, they first make some assumptions on noise, e.g., having a sparse representation. Then the corrupted target data are iteratively recovered to avoid negative transfer. However, real-world noise is often severe and unpredictable, which generally degrades the performance of the above assumption based methods. Even worse, existing works in cross-domain reconstruction are always optimized by the Augmented Lagrange Multiplier (ALM) technique (Lin, Chen, and Ma 2010), resulting in the absence of convergence guarantee. Therefore, effectively learning a robust transformation for domain adaptation is still challenging as the target data may be contaminated by various kinds of complex noise in real-world applications (e.g., contiguous occlusions and shadows exist simultaneously on images).

Recent studies derived from information-theoretic learning (Príncipe, Xu, and Fisher 2000) have shown their superiority in robust learning (Chen et al. 2016; 2017b). In particular, based on Renyi's quadratic entropy, correntropy is proposed and proven to have the theoretical foundation of handling unpredictable noise and outliers (Liu, Pokharel, and Príncipe 2007). In conventional supervised learning (e.g., dimensionality reduction (Yuan and Hu 2009) and classification (He, Zheng, and Hu 2011)), the correntropy based objectives provide strong robustness. However, these objectives fail to attain domain alignment and robustness simultaneously. Without knowledge transfer, the learning machines trained from source domain are invalid for target domain.

In this paper, we make great efforts to address the challenges discussed above and propose a novel reconstruction based method in an information-theoretic setting, named Robust Information-Theoretic Domain Adaptation (RIDA). To robustly uniform the geometrical properties of two domains, RIDA aims at transforming all data into a new space

*Corresponding Author.

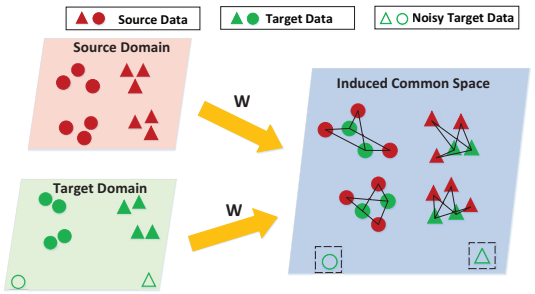


Figure 1: Overview of RIDA. The circle and the triangle denote two different classes. RIDA transforms all data into a new space by the matrix $\mathbf{W}$ based on reconstruction. In the process, the contaminated target points (i.e., outliers) are detected and removed. Thus, each clean target point can be accurately reconstructed by its neighbors in source domain.

such that: the contaminated target data are detected and neglected; the clean target points that lie in a specific neighborhood can be locally reconstructed by the corresponding source-domain neighborhood. In this term, the transformation is supposed to minimize the marginal distribution difference between domains and transfer the discriminative information from source domain to assist the target-domain recognition tasks. For this purpose, we learn a transformation matrix and a reconstruction coefficients matrix through a single optimization and the following three terms compose the objective function. (1) Correntropy is explored to minimize the sample-specific error between the target samples and their reconstructions by the source samples. Benefiting from exploring this robust measurement, RIDA is distinguished with putting more emphasis on the clean target samples and eliminating the negative influence of the contaminated samples on cross-domain reconstruction. (2) The l_1-norm constraint is imposed on the coefficients matrix to capture the intrinsic relatedness of two domains, which ensures the neighborhood-to-neighborhood reconstruction. (3) RIDA utilizes a relative entropy based regularization for the transformation matrix to avoid some trivial solutions and reap the theoretic advantages (i.e., non-negativity and scale-invariance). The overview of RIDA is illustrated in Figure 1. In summary, the contribution of this paper is three-fold.

- RIDA is of great robustness in knowledge transfer by removing the contaminated data without any specific assumptions on noise. It is one of the early solutions for the challenging problems where the target data are contaminated even by different kinds of severe noise.
- We develop an effective half-quadratic technique for optimization, which simplifies the non-convex information-theoretic loss function to the quadratic problems. Moreover, the convergence proof is provided.
- RIDA brings considerable improvements in cross-domain object recognition and face recognition, compared with the state-of-the-art domain adaptation methods.

Related Works

Domain Adaptation

Significant efforts in domain adaptation have been spent on feature extraction learning which transfers both domains to a common subspace where the distributions of two domains are approximately identical (Gong et al. 2012; Long et al. 2013; Zhou et al. 2014; Wang et al. 2015). In this category, cross-domain data reconstruction has been well studied to uncover data relatedness especially when the data from two domains are drawn from a union of multiple subspaces (Shao, Kit, and Fu 2014; Jhuo et al. 2012; Ding, Shao, and Fu 2015; Xu et al. 2016; Zhang, Zuo, and Zhang 2016). To further enhance the robustness, most of the reconstruction methods introduce an error matrix $\mathbf{E}$ and impose a norm constraint on $\mathbf{E}$ (e.g., l_1-norm and $l_{2,1}$-norm) based on a certain assumption about noise (e.g., sparsity). Afterwards, they iteratively correct $\mathbf{E}$ and recover contaminated data from $\mathbf{E}$. This noise correction strategy is empirically validated to be effective for arbitrarily sparse noise (Wright et al. 2009). However, real-world noise is often unpredictable. Recent theoretical analysis and experimental results verify that a certain noise assumption cannot deal with complex noise very well (e.g., large contiguous occlusions), especially when the data are simultaneously corrupted by several types of noise (He et al. 2015). Moreover, these cross-domain reconstruction based methods are hampered by the lack of convergence guarantee with the ALM based optimization.

Correntropy

In information-theoretic learning, correntropy is defined as a similarity measure between random variables X and Y:

$$V(X,Y) = E[k(X-Y)], \tag{1}$$

where $k(.)$ is a kernel function and $E[\cdot]$ denotes the expectation operator. Correntropy is closely related with Welsch M-estimators (Huber 2011) and has robust theoretical foundation (Liu, Pokharel, and Príncipe 2007; Chen et al. 2017a). Moreover, correntropy owns the properties of symmetry, positivity and boundedness. Given a finite number of samples $\{(x_i, y_i)\}_{i=1}^n$, Eq. (1) is extended to the following empirical measure, named Correntropy Induced Metric (CIM):

$$CIM(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(k(0) - k(x_i - y_i)). \tag{2}$$

The value of CIM is mainly decided by the kernel function along the line $X = Y$ (Liu, Pokharel, and Príncipe 2007). CIM has been successfully applied to many supervised learning machines and is proven to be applicable under a variety of unpredictable noisy environments (e.g., missed entries, incorrect labeling and dense corruptions) (Yuan and Hu 2009; He, Zheng, and Hu 2011). However, how to exploit the robust value of the correntropy to address the domain adaptation problems needs further investigation.

Method

In this section, we introduce the Robust Information-Theoretic Domain Adaptation (RIDA) algorithm in detail.

Problem Formulation

Let $\mathbf{X}_t = [\mathbf{x}_1^t, ..., \mathbf{x}_m^t] \in \mathbb{R}^{d\times m}$ be the target data matrix consisting m unlabeled samples from a target domain. Note that a large portion of the target data are seriously contaminated under severe noise conditions. Let $\mathbf{X}_s = [\mathbf{x}_1^s, ..., \mathbf{x}_n^s] \in \mathbb{R}^{d\times n}$ be the source data matrix consisting n labeled samples from a source domain. The corresponding label matrix is denoted as $\mathbf{Y}_s = [y_1^s, ..., y_n^s] \in \mathbb{R}^n$. We denote $P_t(\mathbf{X}_t)$ and $P_s(\mathbf{X}_s)$ as the marginal probability distributions of $\mathbf{X}_t$ and $\mathbf{X}_s$ respectively, $P_t(\mathbf{X}_t) \neq P_s(\mathbf{X}_s)$. The task of RIDA is to eliminate the negative impact of the contaminated target samples during learning a shared transformation $\mathbf{W}$ such that $P_t(\mathbf{W}\mathbf{X}_t)$ and $P_s(\mathbf{W}\mathbf{X}_s)$ will be approximately equal for successful knowledge transfer.

In the following sections, we start learning a transformation by considering the basic case without noise contamination. Afterwards, the loss function is extended to severe noise conditions, which gives the formulation of our RIDA.

The Basic Condition without Noise Contamination In this basic case, we aim at transforming the target and the source data into a common space by reconstructing all the target data, where the reconstruction should reveal the data structures of two domains. Specifically, it is expected that in the new space, the target points in a neighborhood can be reconstructed by the corresponding neighborhood in the source domain. In other word, the transformed samples from two domains have similar geometrical properties. As a result, the distributions of two domains are approximately identical in the new space.

In order to achieve the above goal, we learn a transformation matrix and a reconstruction coefficients matrix through a single optimization by: 1) minimizing reconstruction error; 2) finding sparse reconstruction coefficients; 3) regularizing the transformation. Mathematically, the problem can be formulated as the following function:

$$\min_{\mathbf{W},\mathbf{C}} \|\mathbf{W}\mathbf{X}_t - \mathbf{W}\mathbf{X}_s\mathbf{C}\|_F^2 + \lambda_1\|\mathbf{C}\|_1 + \lambda_2 R(\mathbf{W}), \tag{3}$$

where $\mathbf{W} \in \mathcal{R}^{d\times d}$ is the transformation matrix, $\mathbf{C} \in \mathcal{R}^{n\times m}$ is the reconstruction coefficients matrix, $R(\mathbf{W})$ is the regularizer of $\mathbf{W}$, λ_1 and λ_2 are the trade-off parameters. The rationality of the loss function in Eq. (3) is illustrated as follows. First, the l_1-norm constraint on $\mathbf{C}$ (i.e., $\|\mathbf{C}\|_1 = \sum_{i,j}|\mathbf{C}_{ij}|$) assures the sparsity. The sparse coefficients help in uncovering the relatedness and ensuring the neighborhood-to-neighborhood reconstruction. Second, the regularizer is integrated to avoid some trivial solutions (e.g., zero matrix) and control the complexity of $\mathbf{W}$.

The Severe Noise Conditions In practical and general domain adaptation scenarios, the target data are often contaminated by unpredictable noise, especially when it comes from the web. In these noise conditions, the target data consist of two parts: the contaminated target points (i.e., the outliers) and the uncontaminated target points (i.e., the clean samples). The contaminated target points are far from the clean data and have no corresponding neighbourhood in the source domain. However, the measurement of reconstruction error in Eq. (3) (i.e., the Frobenius norm) is very sensitive to the noise or outliers, which puts more emphasis on the contaminated points. Consequently, based on the learnt $\mathbf{W}$, the clean target points cannot be accurately represented by the corresponding source samples, leading to inaccurate data alignment and significantly degraded prediction performance.

To alleviate the influence of noise and improve the robust adaptation ability, the key challenge is how to detect the contaminated target points and measure the reconstruction error without their interference. We investigate the theoretical foundation of correntropy in robustness against complex noise and originally introduce it into cross-domain data reconstruction. To this end, the CIM in Eq. (2) with Gaussian kernel $k(x) = exp(-x^2/\sigma^2)$ is explored to be the information-theoretic measurement and the following domain adaptation formulation is obtained:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{C}} \ & \sum_{i=1}^{m}\{1 - exp(-\|(\mathbf{W}\mathbf{X}_t - \mathbf{W}\mathbf{X}_s\mathbf{C})^i\|_2^2/\sigma^2)\} \\ & + \lambda_1\|\mathbf{C}\|_1 + \lambda_2 R(\mathbf{W}),\end{aligned} \tag{4}$$

where $(\mathbf{W}\mathbf{X}_t - \mathbf{W}\mathbf{X}_s\mathbf{C})^i$ denotes the i-th column of the error $(\mathbf{W}\mathbf{X}_t - \mathbf{W}\mathbf{X}_s\mathbf{C})$. This column-wise minimization penalizes the error corresponding to a single sample as a whole. It is derived from the group sparsity and is used to control sample-specific error.

The Regularizer $R(\mathbf{W})$ Regarding the regularizer $R(\mathbf{W})$, we consider the Mahalanobis distance matrix $\mathbf{A}$ ($\mathbf{A} = \mathbf{W}^T\mathbf{W}$) and require the regularizer to be able to reduce the distance between $\mathbf{A}$ and the identity matrix $\mathbf{I}$. This distance based regularization leads to satisfactory experimental results in practical domain adaptation problems (Hoffman et al. 2014). Specifically, $R(\mathbf{W})$ employs the relative entropy based measurement proposed in (Davis et al. 2007) to define the distance between $\mathbf{A}$ and $\mathbf{I}$, as:

$$R(\mathbf{W}) = D(\mathbf{A}, \mathbf{I}) = tr(\mathbf{A}\mathbf{I}^{-1}) - log\ det(\mathbf{A}\mathbf{I}^{-1}). \tag{5}$$

$R(\mathbf{W})$ in Eq. (5) reaps the theoretic advantages: 1) non-negativity and 2) scale-invariance with an invertible linear transformation $\mathbf{S}$, i.e., $D(\mathbf{S}^T\mathbf{A}\mathbf{S}, \mathbf{S}^T\mathbf{I}\mathbf{S}) = D(\mathbf{A}, \mathbf{I})$.

The Formulation of RIDA Substituting Eq. (5) into Eq. (4), we derive a novel method named Robust Information-Theoretic Domain Adaptation (RIDA) to address the challenging domain adaptation problems under severe noise conditions. It inherits the robustness by minimizing the information-theoretic objectives $F(\mathbf{W}, \mathbf{C})$:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{C}} \ & F(\mathbf{W},\mathbf{C}) = \sum_{i=1}^{m}\{1 - exp(-\|(\mathbf{W}\mathbf{X}_t - \mathbf{W}\mathbf{X}_s\mathbf{C})^i\|_2^2/\sigma^2)\} \\ & + \lambda_1(tr(\mathbf{W}^T\mathbf{W}) - log\ det(\mathbf{W}^T\mathbf{W})) + \lambda_2\|\mathbf{C}\|_1.\end{aligned} \tag{6}$$

Once we solve the problem in Eq. (6) and obtain the optimal $\mathbf{W}$, we can transform all data into a common space using $\mathbf{W}$. Finally, in this common space, the predictive models trained from the labeled source data can be directly applied to the target domain with high-confidence predictions.

Discussion: In contrast to the previous domain adaptation methods which recover corrupted target data from error matrix, our proposed RIDA is based on information-theoretic metric and has a clear theoretical foundation of robustness.

It can effectively deal with the conditions: a mass of target points are corrupted; the corruptions are dense; and the corruptions are caused by several kinds of severe noise. Specifically, based on the local property of the correntropy, RIDA treats each target sample adaptively during cross-domain reconstruction. The corrupted samples, which have no neighbors in the source domain, are prone to large reconstruction error. Hence, they have relatively stable correntropy values and only make limited impacts on the minimization. That is, RIDA removes the contaminated points (i.e., their error is greatly suppressed), and mainly uses the clean target samples to learn the transformation and the coefficients matrix. In this way, RIDA accurately reduces the distribution difference for knowledge transfer without any specific assumptions on noise.

Optimization

Since Eq. (6) is non-linear and non-convex, it is difficult to be directly optimized. In this section, we explore the Half-Quadratic (HQ) technique (Geman and Reynolds 1992) and propose an efficient procedure to iteratively optimize the augmented function of RIDA in an enlarged parameter space. Different from the ALM based optimization always applied in the existing cross-domain representation methods, our proposed procedure has the provable convergence guarantee due to the theoretic properties of our information-theoretic loss function.

Conjugate Function and HQ Form Based on the theory of convex conjugate functions (Boyd and Vandenberghe 2004; He et al. 2015), the following proposition can be derived, which enables Eq. (6) to be minimized in HQ way:

Proposition 1. *There exists a convex conjugated function* $\varphi(p)$ *of* $g(x) = 1 - exp(-x^2/\sigma^2)$*, such that:*

$$g(x) = \min_{p \in R}(p\|x\|^2 - \varphi(p)), \tag{7}$$

and for a fixed x*, the minimization is reached at* $p = exp(-x^2/\sigma^2)$.

According to Proposition 1, the first term of $F(\mathbf{W}, \mathbf{C})$ in Eq. (6) can be translated to the following form:

$$\begin{aligned} &\sum_{i=1}^{m}\{1 - exp(-\|(\mathbf{W}\mathbf{X}_t - \mathbf{W}\mathbf{X}_s\mathbf{C})^i\|_2^2/\sigma^2)\} \\ =&\min_{p_i}\sum_{i=1}^{m}\{p_i\|(\mathbf{W}\mathbf{X}_t - \mathbf{W}\mathbf{X}_s\mathbf{C})^i\|_2^2 + \varphi(p_i)\}. \end{aligned} \tag{8}$$

The above expression is a basic form in HQ analysis (Geman and Reynolds 1992) where p_i is called the auxiliary variable.

Alterative Minimization based on HQ Analysis Substituting the HQ format of Eq. (8) into Eq. (6), the following augmented objective function of RIDA (i.e., $J(\mathbf{P}, \mathbf{W}, \mathbf{C})$) is obtained in an enlarged parameter space:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{C}} F(\mathbf{W}, \mathbf{C}) = \min_{\mathbf{P},\mathbf{W},\mathbf{C}} J(\mathbf{P}, \mathbf{W}, \mathbf{C}) \\ &= \sum_{i=1}^{m}\{p_i\|(\mathbf{W}\mathbf{X}_t - \mathbf{W}\mathbf{X}_s\mathbf{C})^i\|_2^2 + \varphi(p_i)\} \\ &+ \lambda_1\|\mathbf{C}\|_1 + \lambda_2(tr(\mathbf{W}^T\mathbf{W}) - log\ det(\mathbf{W}^T\mathbf{W})), \end{aligned} \tag{9}$$

where $\mathbf{P} \in R^{m\times m}$ is a diagonal matrix with $\mathbf{P}(i,i) = p_i$.

In HQ analysis, $J(\mathbf{P}, \mathbf{W}, \mathbf{C})$ can be minimized by an alternative strategy which updates one variable with the others fixed. We emphasize that each update is a convex optimization problem in our alternative strategy as follows, which is numerically tractable.

1) Update $\mathbf{P}$. Based on Proposition 1, p_i can be easily updated as:

$$p_i^{r+1} = exp(-\|(\mathbf{W}^r\mathbf{X}_t - \mathbf{W}^r\mathbf{X}_s\mathbf{C}^r)^i\|_2^2/\sigma^2), \tag{10}$$

where $i = 1, ..., m$ and r is the iteration number.

2) Update $\mathbf{W}$. When p_i is fixed, $\varphi(p_i)$ in Eq. (9) becomes a constant and can be removed. In this term, $\mathbf{W}$ is updated by solving the following problem:

$$\begin{aligned} \mathbf{W}^{r+1} =& arg\min_{\mathbf{W}}\sum_{i=1}^{m} p_i^{r+1}\|(\mathbf{W}\mathbf{X}_t - \mathbf{W}\mathbf{X}_s\mathbf{C}^r)^i\|_2^2 \\ &+ \lambda_2(tr(\mathbf{W}^T\mathbf{W}) - log\ det(\mathbf{W}^T\mathbf{W})). \end{aligned} \tag{11}$$

The above loss function is not convex with respect to $\mathbf{W}$. Therefore, we work in terms of the new variable $\mathbf{A} = \mathbf{W}^T\mathbf{W}$. With this alternative definition of variable $\mathbf{A}$, Eq. (11) can be rewritten as a convex optimization problem:

$$\begin{aligned} \mathbf{A}^{r+1} =& arg\min_{\mathbf{A}} Tr(\mathbf{A}(\mathbf{X}_t - \mathbf{X}_s\mathbf{C}^r)\mathbf{P}^{r+1}(\mathbf{X}_t - \mathbf{X}_s\mathbf{C}^r)^T) \\ &+ \lambda_2(tr(\mathbf{A}) - log\ det(\mathbf{A})). \end{aligned} \tag{12}$$

By setting the derivative to zero, the closed-form solution of $\mathbf{A}^{r+1}$ can be obtained:

$$\mathbf{A}^{r+1} = (\mathbf{I} + \frac{1}{\lambda_2}(\mathbf{X}_t - \mathbf{X}_s\mathbf{C}^r)\mathbf{P}^{r+1}(\mathbf{X}_t - \mathbf{X}_s\mathbf{C}^r)^T)^{-1}. \tag{13}$$

Based on the solution of $\mathbf{A}^{r+1}$, the optimal $\mathbf{W}^{r+1}$ can be obtained using eigen-decomposition: $\mathbf{W}^{r+1} = diag(\sqrt{\theta_1^{r+1}}, \ldots, \sqrt{\theta_d^{r+1}})[\mathbf{u}_1^{r+1}, \ldots, \mathbf{u}_d^{r+1}]^T$, where θ_i^{r+1} and $\mathbf{u}_i^{r+1}$ are the i-th eigenvalue and eigenvector of $\mathbf{A}^{r+1}$ respectively.

3) Update $\mathbf{C}$. For minimizing $\mathbf{C}$, the problem in Eq. (9) can be written as follows by fixing $\mathbf{P}$ and $\mathbf{W}$:

$$\begin{aligned} \mathbf{C}^{r+1} =& arg\min_{\mathbf{C}} \lambda_1\|\mathbf{C}\|_1 + \sum_{i=1}^{m} p_i^{r+1}\|(\mathbf{W}^{r+1}\mathbf{X}_t - \mathbf{W}^{r+1}\mathbf{X}_s\mathbf{C})^i\|_2^2 \\ =& arg\min_{\mathbf{C}}\sum_{i=1}^{m}\{\lambda_1\|\mathbf{C}^i\|_1 + \|(\mathbf{U}^{r+1}\mathbf{C}^i - \mathbf{V}^{r+1,i})\|_2^2\}, \end{aligned} \tag{14}$$

where $\mathbf{C}^i$ and $\mathbf{V}^{r+1,i}$ denote the i-th columns of $\mathbf{C}$ and $\mathbf{V}^{r+1}$ respectively, $\mathbf{U}^{r+1} = \sqrt{p_i^{r+1}}\mathbf{W}^{r+1}\mathbf{X}_s$ and $\mathbf{V}^{r+1} = \sqrt{p_i^{r+1}}\mathbf{W}^{r+1}\mathbf{X}_t$. Note that each $\mathbf{C}^i$ in Eq. (14) can be independently solved, which is a standard convex problem in l_1 minimization. Many iterative techniques have been proposed to solve this l_1 minimization. Among them, we employ the feature-sign search technique (Lee et al. 2007) to find the optimal $\mathbf{C}^i$, which has the step-down character that each iteration of the minimization can reduce the objective function. To further save the operation time, just a local optimum of each $\mathbf{C}^i$ is needed in our procedure instead of finding the global solution.

Algorithm 1 Robust Information-Theoretic Domain Adaptation (RIDA)

Require: $\mathbf{X}_s$, $\mathbf{X}_t$, σ, λ_1 and λ_2.
Ensure: $\mathbf{P}$, $\mathbf{W}$, $\mathbf{C}$.
1: $r = 0$, $\mathbf{W}_0 = \mathbf{I}$, $\mathbf{C}_0 = \mathbf{0}$;
2: **while** not convergence **do**
3: Solve $\mathbf{P}^{r+1}$ based on Eq. (10);
4: Solve $\mathbf{W}^{r+1}$ based on Eq. (13);
5: Solve $\mathbf{C}^{r+1}$ by optimizing the problem in Eq. (14);
6: r=r+1.
7: **end while**
8: **Return** $\mathbf{P}$, $\mathbf{W}$ and $\mathbf{C}$.

Overall, the procedure of optimizing the problem in Eq. (9) is summarized in Algorithm 1.

Robustness Explanation and Complexity Analysis The optimization process in Algorithm 1 gives a clear explanation of the robustness in RIDA. Specifically, the value of p_i is small for the contaminated target data due to the properties of correntropy. Thus, the negative influence of the contaminated data is dismissed during updating $\mathbf{W}$ and $\mathbf{C}$, where p_i acts as the weight.

Next, we analyze the complexity of RIDA. As for step 3 and step 4 in Algorithm 1, their computation complexities are $O(dmn+d^2(m+n))$ and $O(d^3+dmn+d^2m+dm^2)$ respectively. Step 5 costs nearly $O(dmn+\sum_i dN_i^2+\sum_i N_i^3)$, where N_i is the number of nonzero entries of $\mathbf{C}^i$. Since $N_i \ll \min(m,n)$, the overall complexity of Algorithm 1 is $O(R(dmn + d^2(m+n) + d^3 + dm^2))$, where R is the number of iteration.

Convergence Proof The following proposition proves that the RIDA updating sequences will converge.

Proposition 2. *The updating scheme in Algorithm 1 converges to the local minimum of* $J(\mathbf{P}, \mathbf{W}, \mathbf{C})$.

Proof. According to Proposition 1, the closed-form solution in Eq. (13) and the step-down character of the feature-sign search technique, we can achieve the following sequence: $J(\mathbf{P}^{r+1}, \mathbf{W}^{r+1}, \mathbf{C}^{r+1}) \leq J(\mathbf{P}^{r}, \mathbf{W}^{r+1}, \mathbf{C}^{r+1}) \leq J(\mathbf{P}^{r}, \mathbf{W}^{r}, \mathbf{C}^{r+1}) \leq J(\mathbf{P}^{r}, \mathbf{W}^{r}, \mathbf{C}^{r})$. That is, the loss function in Eq. (9) is non-increasing in each iteration. Furthermore, according to the properties of correntropy and relative entropy, each term of $F(\mathbf{W}, \mathbf{C})$ in Eq. (6) is bounded downwards. Therefore, $J(\mathbf{W}, \mathbf{C}, \mathbf{P})$ is bounded downwards as well based on the first line in Eq. (9). Hence, the updating scheme in Algorithm 1 ensures to converge. □

Experiments

In this section, we evaluate the proposed method in two domain adaptation related applications: 1) object recognition and 2) face recognition.

Data Preparation

COIL-20 (Nene et al. 1996) dataset contains 1,440 images from 20 objects. The images of objects are taken at pose intervals of 5 degrees, leading to 72 poses per object. Some example images are shown in Figure 2(a), where each image has the resolution of 32×32 pixels with 256 gray levels per pixel. Following the previous work (Long et al. 2013; Xu et al. 2016), the dataset is partitioned into COIL1 (containing all images in the directions $[0°, 85°] \cup [180°, 265°]$) and COIL2 (containing all images in the directions $[90°, 175°] \cup [270°, 355°]$). Consequently, COIL1 and COIL2 have related but different distributions, and two cross-domain datasets are constructed: 1) C1 *vs* C2: the source dataset is COIL1 and the target dataset is COIL2; 2) C2 *vs* C1: the source/target pair in C1 *vs* C2 is switched.

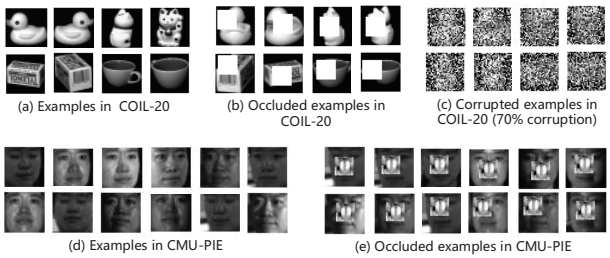


Figure 2: Illustrative images used in the experiments.

CMU-PIE (Sim, Baker, and Bsat 2002) is a benchmark face dataset which includes 41,368 face images from 68 individuals with different poses, illuminations and facial expressions. Each face image is in 256 gray scales per pixel with the size of 32×32. Figure 2(d) shows some example face images. CMU-PIE can be divided into five subsets according to different poses: PIE1 (left pose), PIE2 (upward pose), PIE3 (down pose), PIE4 (front pose), PIE (right pose). As in (Long et al. 2013; Xu et al. 2016), each subset is regarded as a domain and 20 cross-domain face datasets are constructed, i.e., P1 *vs* P2, P1 *vs* P3,..., P5 *vs* P3, P5 *vs* P4.

For the cross-domain datasets above, different types of complex noise is added on a large proportion of the target data to evaluate the robust adaptation ability. Specifically, contiguous occlusions are simulated on COIL-20 such that a random local region (16×16) of the image is replaced by a white square (Fidler, Skocaj, and Leonardis 2006) (see Figure 2(b)). As another kind of noise, the images of COIL-20 are corrupted by randomly replacing 70 percent of pixels with i.i.d samples from a uniform distribution (He, Zheng, and Hu 2011) (see Figure 2(c)). The original images from PIE may have shadows due to different light conditions, we further add contiguous occlusions by randomly replacing a region (16×16) with an unrelated monkey image (Wright et al. 2009) (see Figure 2(e)).

Comparison Methods

We systematically compare the proposed method RIDA with the following state-of-the-art transformation based domain adaption methods: 1) Geodesic Flow Kernel (GFK) (Gong et al. 2012); 2) Transferred Fisher's Linear Discriminant Analysis (TrFLDA) (Si, Tao, and Geng 2010); 3) Joint Distribution Adaptation (JDA) (Long et al. 2013); 4) Latent Sparse Domain Transfer Learning (LSDT) (Zhang, Zuo, and Zhang 2016) and 5) Discriminative Transfer Subspace Learning (DTSL) (Xu et al. 2016). Furthermore, 1-Nearest Neighbor

Table 1: Classification accuracy (%) on the original COIL.

Dataset	Standard Learning		Transfer Learning					
	NN	CESR	GFK	TrFLDA	JDA	LSDT	DTSL	RIDA
C1 *vs* C2	83.61	84.86	85.97	86.67	**89.31**	85.69	88.06	89.17
C2 *vs* C1	82.78	84.69	85.14	85.56	88.47	84.72	**89.17**	88.06

(1-NN) classifier and Correntropy-Based Sparse Representation (CESR) classifier (He, Zheng, and Hu 2011) are also compared as non-transfer baselines. All the transformation based domain adaption methods use 1-NN as the base classifier without parameters tuning. We have tried our best to empirically search the parameter spaces of these six comparison methods to obtain the best results on all the datasets. Our method involves three parameters: λ_1, λ_2 and σ. Across the experiments, we set these parameters by searching the values in the range $[10^{-3}, 10^{3}]$. In general, our method is found to be robust to these parameters.

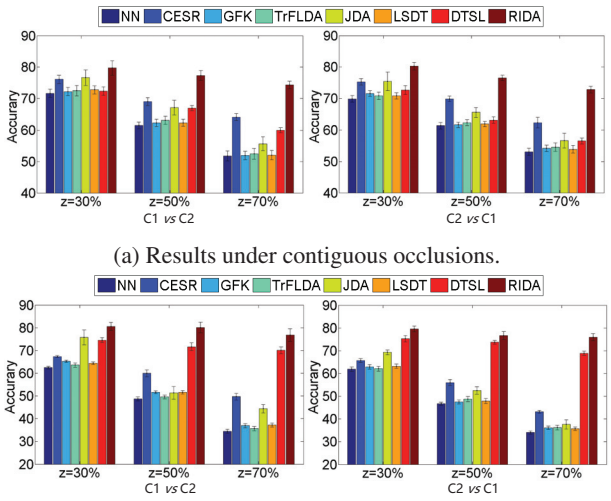
Experimental Results on COIL-20

In this section, we evaluate all the methods under three conditions successively: no noise; a single noise of contiguous occlusions and another single noise of corruptions.

Original Datasets The classification results are shown in Table 1. As can be seen, most of the domain adaptation methods outperform 1-NN and CESR, showing the advantage of information transfer. The results of JDA, DTSL and RIDA are similar and comparable on these two datasets, which are much better than those of other transfer methods.

Contiguous Occlusions with a White Square The source data are composed of all the clean samples in the source domain. For the target data, we randomly pollute z percent of the data using the contiguous occlusions shown in Figure 2(b). The z is set to be different large values (i.e., $z = 30\%, 50\%, 70\%$) to explore the influence of the ratio of the contaminated data. The experiments are randomly repeated 10 times and the average classification results on the unlabeled target domain are shown in Figure 3(a). The following observations can be drawn. (1) The performance of existing domain adaptation methods (i.e., GFK, TrFLDA, JDA, LSDT and DTSL) is greatly degraded due to the added occlusions and is worse than CESR and RIDA. It demonstrates that large occlusions deteriorate the performance of domain adaptation while the correntropy is a more effective similarity measure to deal with these occlusions. (2) Putting emphasis on reducing the distribution difference, our method successfully transfers information across domains and consistently provides much higher accuracy (up to 10% improvement) than the traditional classifier CESR across all the datasets and the numbers of z. (3) When the number z increases, the performance of all the other comparison methods decreases rapidly. By contrast, the accuracy of RIDA decreases more slowly, showing its advantage of handling massive noise pollution.

Random Pixel Corruptions Following the similar setting as before, the source domain is uncontaminated and z percent of the target data are randomly corrupted as shown in Figure 2(c). The average classification results over 10 random repetitions across different numbers of z are shown in Figure 3(b). Some observations can be concluded. First, as can be seen from Figure 2(c), the corrupted images from target data are even barely recognized to the human eye, indicating greater distribution difference between the domains than that of the images in the above experiments. In this difficult case, the robust domain adaptation methods (i.e., DTSL and RIDA) can transfer the discriminating power from the source domain to the noisy target domain, thus outperform the other non-robust methods and non-transfer methods. Second, our method obtains significant improvement compared with DTSL. The possible reason is that DTSL employs the l_1-norm for error modeling based on the sparse assumption, which has the limited capability of recovering clean data from these severe corruptions. By contrast, our correntropy based method can effectively detect and suppress various kinds of complicated noise. Third, with the increasing number of z, the accuracy of our method is still higher than 75% even when $z = 70\%$.



(a) Results under contiguous occlusions.

(b) Results under random pixel corruptions.

Figure 3: Cross-domain object classification accuracy (%) and standard variation (%) on C1 *vs* C2 and C2 *vs* C1.

Experimental Results on CMU-PIE

On CMU-PIE, we conduct experiments under two different conditions: low-level shadow corruptions caused by varied illuminations on the original images; the combined noise of contiguous occlusions and shadow corruptions.

Original Datasets The classification results are shown in Table 2 and we achieve the following observations. (1) CESR outperforms our method on P2 *vs* P4 and P3 *vs* P4. A possible explanation is that each face with front pose can be approximatively expressed as a linear combination of some faces with upward pose or down pose, due to the much similar distributions. Therefore, the non-transfer and expression based classifier CESR is more applicable. But we would like to mention that our method always performs well on the

Table 2: Classification accuracy (%) on the original CMU-PIE with low-level shadow corruptions.

Dataset	Standard Learning		Transfer Learning					
	NN	CESR	GFK	TrFLDA	JDA	LSDT	DTSL	RIDA
P1 *vs* P2	26.09	44.81	26.15	39.23	58.81	26.21	**65.87**	60.10
P1 *vs* P3	26.59	48.41	27.27	35.48	54.23	26.53	**64.09**	62.32
P1 *vs* P4	30.67	61.07	31.15	51.46	84.50	30.64	**82.03**	75.46
P1 *vs* P5	16.67	27.51	17.59	27.21	49.75	16.91	**54.90**	48.22
P2 *vs* P1	24.49	38.39	25.24	31.36	57.62	24.43	45.04	**58.31**
P2 *vs* P3	46.63	68.75	47.37	33.95	62.93	46.57	53.49	**71.02**
P2 *vs* P4	54.07	**84.53**	54.25	61.67	75.82	54.10	71.43	80.29
P2 *vs* P5	26.53	43.32	27.08	25.12	39.89	26.53	47.94	**52.51**
P3 *vs* P1	21.37	32.80	21.82	40.40	50.96	21.40	52.49	**54.89**
P3 *vs* P2	41.01	56.66	43.16	34.56	57.95	41.07	55.56	**65.25**
P3 *vs* P4	46.53	**82.94**	46.41	66.60	68.45	46.53	77.50	80.84
P3 *vs* P5	26.23	43.50	26.78	37.62	39.95	26.23	54.11	**60.60**
P4 *vs* P1	32.95	50.36	34.24	74.04	80.58	32.89	**81.54**	78.45
P4 *vs* P2	62.68	84.47	62.92	78.45	82.63	62.74	85.39	**88.34**
P4 *vs* P3	73.22	90.38	73.35	78.13	87.25	73.10	82.23	**91.12**
P4 *vs* P5	37.19	57.60	37.38	58.64	54.66	37.38	72.61	**75.12**
P5 *vs* P1	18.49	31.33	20.35	42.74	46.46	18.46	**52.19**	48.56
P5 *vs* P2	24.19	38.37	24.62	38.43	42.05	24.19	49.41	**52.67**
P5 *vs* P3	28.31	49.33	28.49	46.02	53.31	28.31	58.45	**62.01**
P5 *vs* P4	31.24	61.16	31.33	57.49	57.01	31.21	64.31	**68.58**
Mean	34.76	54.78	35.35	47.93	60.24	34.78	63.53	**66.73**

Table 3: Classification accuracy (%) on CMU-PIE under added contiguous occlusions

Dataset	Standard Learning		Transfer Learning					
	NN	CESR	GFK	TrFLDA	JDA	LSDT	DTSL	RIDA
P1 *vs* P2	19.15	37.69	19.28	27.93	29.96	19.28	48.13	**50.03**
P1 *vs* P3	20.53	41.54	20.53	24.63	34.80	20.53	46.20	**49.26**
P1 *vs* P4	24.33	52.60	24.36	37.13	53.17	24.30	65.70	**69.21**
P1 *vs* P5	12.93	21.63	13.24	19.06	24.26	12.81	38.66	**39.15**
P2 *vs* P1	18.52	33.28	18.67	20.89	35.32	18.58	36.40	**43.64**
P2 *vs* P3	30.09	**58.21**	30.21	20.89	35.91	30.09	38.48	52.63
P2 *vs* P4	40.04	**79.81**	40.25	43.77	49.23	40.19	61.97	71.52
P2 *vs* P5	19.49	36.52	19.67	16.36	22.49	19.49	38.05	**42.52**
P3 *vs* P1	16.81	28.24	16.93	26.20	28.93	16.84	38.81	**40.76**
P3 *vs* P2	28.48	47.89	28.55	22.41	31.31	28.42	40.09	**48.99**
P3 *vs* P4	34.09	**75.73**	34.15	46.80	39.53	34.09	63.35	69.39
P3 *vs* P5	19.06	38.24	19.30	23.16	24.75	18.93	39.52	**46.81**
P4 *vs* P1	25.03	43.86	25.27	48.56	52.49	24.94	63.69	**64.89**
P4 *vs* P2	44.14	77.66	44.32	51.26	53.22	44.26	73.05	**78.82**
P4 *vs* P3	51.59	83.64	51.53	52.45	58.09	51.59	70.28	**84.65**
P4 *vs* P5	27.33	51.16	27.82	38.24	34.80	27.45	58.21	**61.15**
P5 *vs* P1	14.26	25.75	14.26	27.55	30.70	14.26	36.42	**38.49**
P5 *vs* P2	18.48	32.97	18.48	25.29	22.77	18.48	40.39	**43.19**
P5 *vs* P3	20.40	42.22	20.65	28.31	27.70	20.59	45.04	**45.53**
P5 *vs* P4	25.20	55.00	25.32	40.01	33.34	25.14	49.05	**55.60**
Mean	25.50	48.18	25.64	32.05	36.14	25.51	49.57	**54.81**

other cross-domain datasets with larger distribution differences. (2) Note that the source data on CMI-PIE are mildly corrupted as well. RIDA and DTSL are shown to outperform the remaining transfer methods, since they can generally reconstruct each uncorrupted target point (or recovered point) by its uncorrupted neighbors in the source domain and transfer the discriminating information accurately. (3) Our method achieves higher accuracy rates than DTSL on 14 datasets. For instance, our method achieves almost 10% improvements compared to DTSL on P2 *vs* P4 and P4 *vs* P3. Moreover, the average performance of our method is much better than all the other competitors. These results illustrate the reliable cross-domain performance of our method.

Contiguous Occlusions with an Unrelated Image We further randomly occlude 50 percent of the number of the target data using the unrelated monkey image shown in Figure 2(e). Contiguous occlusions and shadows may exist simultaneously on these target images, leading to large outliers. The average classification results over 10 random repetitions are shown in Table 3 and the following observations can be concluded. First, with the added occlusions, the accuracy of all the methods decreases, especially JDA. Second, our method outperforms all the other domain adaptation methods in terms of average accuracy. RIDA can assign small weights to the large outliers, and put more emphasis on the uncontaminated points during learning the reconstruction and the transformation. As a result, the learnt transformation can explicitly reduce the distribution difference.

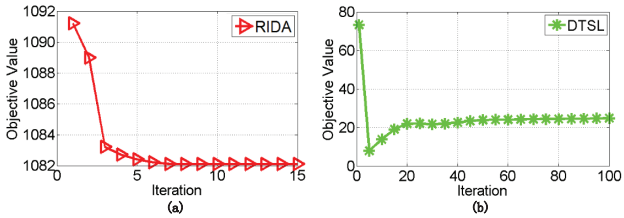


Figure 4: Convergence of RIDA and DTSL on C1 *vs* C2.

Convergence Analysis

The convergence of RIDA has been proven in Section 3. In this section, we experimentally plot its convergence on C1 *vs* C2 in Figure 4(a). For comparison, the convergence of DTSL has also been shown in Figure 4(b). As can be seen, our objective function decreases in every iteration and the optimization process converges after less than 10 iterations. In contrast, the objective value of DTSL is volatile and there is still no convergence after 100 steps of iterations. Similar observations can be drawn from other datasets as well.

Conclusion

In this paper, we have proposed a novel domain adaptation method inspired from correntropy. The key idea is to seek a shared feature space based on cross-domain reconstruction and incorporate the removal of contaminated target data into this seeking process, resulting in an accurate alignment between two domains. Without any specific assumptions on noise, the proposed method achieves its main advantage in the strong robustness for the challenging domain adaptation problems where the target data are contaminated by different kinds of severe and complex noise. Furthermore, an effective half-quadratic technique has been developed, guaranteeing the convergence of RIDA. Comprehensive experimental results validate the effectiveness and the noise suppression ability of the proposed method. In the future, we plan to facilitate the robustness by exploring more knowledge (e.g., class information) from two domains.

Acknowledgments

This work is supported by Natural Science Foundation of China (61502466,61672501,61473289).

References

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Chen, B.; Xing, L.; Xu, B.; Zhao, H.; and Príncipe, J. C. 2016. Insights into the robustness of minimum error entropy estimation. *IEEE transactions on neural networks and learning systems*.

Chen, B.; Liu, X.; Zhao, H.; and Príncipe, J. C. 2017a. Maximum correntropy kalman filter. *Automatica* 76:70–77.

Chen, B.; Xing, L.; Xu, B.; Zhao, H.; Zheng, N.; and Príncipe, J. C. 2017b. Kernel risk-sensitive loss: definition, properties and application to robust adaptive filtering. *IEEE Transactions on Signal Processing* 65(11):2888–2901.

Davis, J. V.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. S. 2007. Information-theoretic metric learning. In *Proceedings of the Twenty-Fourth International Conference on Machine Learning*, 209–216.

Ding, Z.; Shao, M.; and Fu, Y. 2015. Deep low-rank coding for transfer learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3453–3459.

Fidler, S.; Skocaj, D.; and Leonardis, A. 2006. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(3):337–350.

Geman, D., and Reynolds, G. 1992. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14(3):367–383.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the Twenty-Fifth IEEE Conference on Computer Vision and Pattern Recognition*, 2066–2073.

He, R.; Zhang, Y.; Sun, Z.; and Yin, Q. 2015. Robust subspace clustering with complex noise. *IEEE Transactions on Image Processing* 24(11):4001–4013.

He, R.; Zheng, W.-S.; and Hu, B.-G. 2011. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(8):1561–1576.

Hoffman, J.; Rodner, E.; Donahue, J.; Kulis, B.; and Saenko, K. 2014. Asymmetric and category invariant feature transformations for domain adaptation. *International Journal of Computer Vision* 109(1-2):28–41.

Huber, P. J. 2011. Robust statistics. In *International Encyclopedia of Statistical Science*. Springer. 1248–1251.

Jhuo, I.-H.; Liu, D.; Lee, D.; and Chang, S.-F. 2012. Robust visual domain adaptation with low-rank reconstruction. In *Proceedings of the Twenty-Fifth IEEE Conference on Computer Vision and Pattern Recognition*, 2168–2175.

Lee, H.; Battle, A.; Raina, R.; and Ng, A. Y. 2007. Efficient sparse coding algorithms. In *Proceedings of the Twenty-First Annual Conference on Advances in Neural Information Processing Systems*, 801–808.

Lin, Z.; Chen, M.; and Ma, Y. 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*.

Liu, W.; Pokharel, P. P.; and Príncipe, J. C. 2007. Correntropy: Properties and applications in non-gaussian signal processing. *IEEE Transactions on Signal Processing* 55(11):5286–5298.

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2013. Transfer feature learning with joint distribution adaptation. In *Proceedings of the Fourteenth International Conference on Computer Vision*, 2200–2207.

Nene, S. A.; Nayar, S. K.; Murase, H.; et al. 1996. Columbia object image library (coil-20).

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2):199–210.

Príncipe, J. C.; Xu, D.; and Fisher, J. 2000. Information theoretic learning. *Unsupervised adaptive filtering* 1:265–319.

Shao, M.; Kit, D.; and Fu, Y. 2014. Generalized transfer subspace learning through low-rank constraint. *International Journal of Computer Vision* 109(1-2):74–93.

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

Sim, T.; Baker, S.; and Bsat, M. 2002. The cmu pose, illumination, and expression (pie) database. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 53–58.

Wang, W.; Wang, H.; Zhang, C.; and Xu, F. 2015. Transfer feature representation via multiple kernel learning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 3073–3079.

Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2):210–227.

Xu, Y.; Fang, X.; Wu, J.; Li, X.; and Zhang, D. 2016. Discriminative transfer subspace learning via low-rank and sparse representation. *IEEE Transactions on Image Processing* 25(2):850–863.

Yuan, X.-T., and Hu, B.-G. 2009. Robust feature extraction via information theoretic learning. In *Proceedings of the Twenty-Sixth International Conference on Machine Learning*, 1193–1200.

Zhang, L.; Zuo, W.; and Zhang, D. 2016. Lsdt: Latent sparse domain transfer learning for visual adaptation. *IEEE Transactions on Image Processing* 25(3):1177–1191.

Zhou, J. T.; Pan, S. J.; Tsang, I. W.; and Yan, Y. 2014. Hybrid heterogeneous transfer learning through deep learning. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2213–2220.