Robust Transfer Metric Learning for Image Classification

Zhengming Ding, *Student Member, IEEE*, and Yun Fu, *Senior Member, IEEE*

***Abstract*— Metric learning has attracted increasing attention due to its critical role in image analysis and classification. Conventional metric learning always assumes that the training and test data are sampled from the same or similar distribution. However, to build an effective distance metric, we need abundant supervised knowledge (i.e., side/label information), which is generally inaccessible in practice, because of the expensive labeling cost. In this paper, we develop a robust transfer metric learning (RTML) framework to effectively assist the unlabeled target learning by transferring the knowledge from the well-labeled source domain. Specifically, RTML exploits knowledge transfer to mitigate the domain shift in two directions, i.e., sample space and feature space. In the sample space, domain-wise and class-wise adaption schemes are adopted to bridge the gap of marginal and conditional distribution disparities across two domains. In the feature space, our metric is built in a marginalized denoising fashion and low-rank constraint, which make it more robust to tackle noisy data in reality. Furthermore, we design an explicit rank constraint regularizer to replace the rank minimization NP-hard problem to guide the low-rank metric learning. Experimental results on several standard benchmarks demonstrate the effectiveness of our proposed RTML by comparing it with the state-of-the-art transfer learning and metric learning algorithms.**

***Index Terms*— Transfer learning, cross-domain metric, marginalized denoising.**

I. Introduction

METRIC learning plays a fundamental role in image analysis and pattern recognition fields, which has been demonstrated that effective distance metrics built on large labeled training data could greatly facilitate the recognition performance for test data [1]–[3]. Conventional metric learning methods always achieve promising results when there are sufficient labeled training data [1], [2], [4], [5]. The training strategy is usually to minimize the distance between two samples with the same label, and otherwise maximize that of between-class samples. Generally, distance metric learning methods tend to transform the original data to a new space with the metric, which can be split into two fashions based on whether the geometry structure is incorporated or not. Specifically, global metric learning methods manage to pull all the data points with the same class label close together for compactness while pushing those from different classes far apart for separability [3]–[5], whilst local metric learning methods are designed to preserve the geometry structure of data with the label information [6]–[9].

However, we need a lot of supervised information to build an effective distance metric, which is generally not available in many real-world applications [10]. Actually, we always confront the scenario that no or limited labeled data in the target domain are available during the training stage. Indeed, distance metrics learned only in a well-labeled source domain cannot be directly reused in the target domain, because the significant distribution divergence across two domains is not explicitly taken into consideration. This domain shift would lead to the metrics trained in source domain invalid in target domain. On the other hand, it is very expensive to manually annotate the unlabeled data. Therefore, it is essential to reuse well-labeled source data and reduce the domain shift through distance metric learning.

Recently, transfer learning [10] has been verified as an appealing approach to deal with such challenges by making use of the well-learned knowledge from external source datasets. Recent research efforts on transfer learning have shown attractive results by learning a common space in which source knowledge can be adapted to facilitate the target learning [11]–[14]. Therefore, it is very important to capture the rich and useful knowledge across two domains through learning a valid metric [15]–[19]. However, current cross-domain metric learning algorithms only consider the marginal distribution difference across two domains, while ignoring the conditional distribution disparity between them. Moreover, they mainly focus on one direction alignment (sample space) to guide the knowledge transfer, without considering the feature space alignment. Hence, they are not effective enough in dealing with unlabeled target learning.

In this paper, we propose a novel Robust Transfer Metric Learning (RTML) algorithm for cross-domain image classification (Figure 1). The core idea of RTML is to seek a robust transfer low-rank metric to address the marginal and conditional distribution differences across two domains. In this way, our cross-domain metric could well adapt well-learned source knowledge to facilitate the target learning. Furthermore, we design a marginalized denoising scheme to seek a more robust cross-domain metric to real-world images. To this end, we seek a cross-domain metric through aligning source

Manuscript received April 6, 2016; revised September 30, 2016 and November 6, 2016; accepted November 15, 2016. Date of publication November 22, 2016; date of current version December 6, 2016. This work was supported in part by NSF IIS under Award 1651902, in part by NSF CNS under Award 1314484, in part by ONR under Award N00014-12-1-1028, in part by the ONR Young Investigator Award under Grant N00014-14-1-0484, and in part by the U.S. Army Research Office Young Investigator Award under Grant W911NF-14-1-0218. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vladimir Stankovic.

Z. Ding is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: allanding@ece.neu.edu).

Y. Fu is with the Department of Electrical and Computer Engineering, College of Computer and Information Science, Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2631887

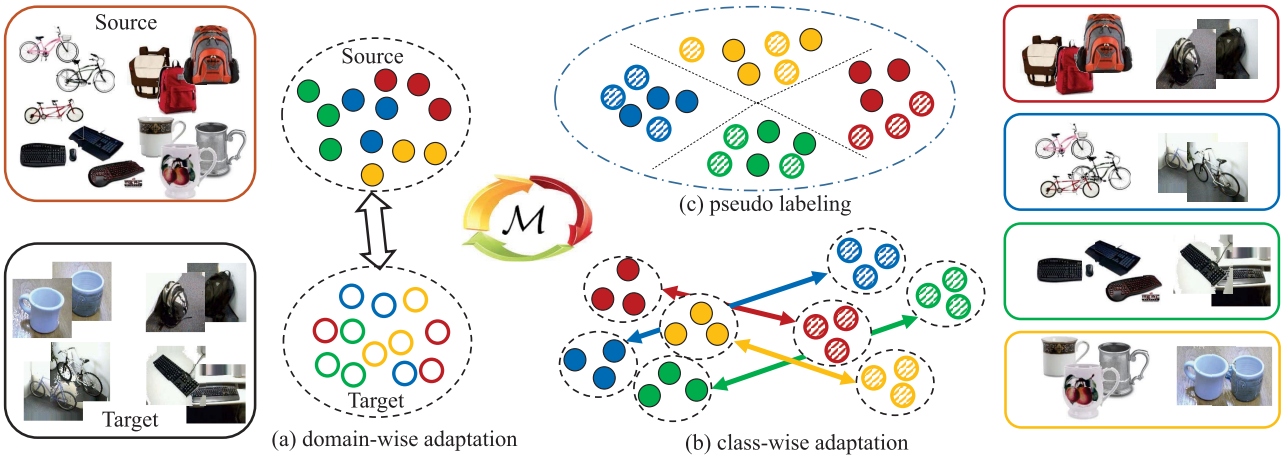


Fig. 1. Illustration of our proposed Robust Transfer Metric Learning (RTML). Note that the same color means the same class (Here we have 4 classes). There are two domains with different distributions in the original space. RTML aims to build a cross-domain metric to mitigate the domain shift. First, we propose domain-/class-wise adaptation (a,b) strategy to alleviate the disparity of the marginal and conditional distributions between two domains. Moreover, the class-wise adaptation term is iteratively optimized through pseudo labels of target domain (c). Furthermore, we develop marginalized denoising fashion and low-rank constraint to seek a robust and effective cross-domain metric.

and target in two directions, i.e., sample space and feature space.

A. Our Contributions

We aim to seek a robust cross-domain metric to boost the unlabeled target learning. Hence, we summarize our contributions in the three folds as follows:

- First of all, we propose an iterative refine manner to optimize the class-wise adaptation term to uncover more intrinsic source knowledge for target learning. Hence, marginal and conditional distribution differences across two domains are both leveraged through a shared cross-domain metric.
- Secondly, marginalized denoising scheme is developed to improve the robustness of the learned metric. Specifically, our metric works in a denoising auto-encoder data reconstruction way, aiming to recover the clean knowledge from the manually corrupted data.
- Simultaneously, low-rank constraint is incorporated to guide the cross-domain metric learning by uncovering more common feature structure across two domains. To recover a better low-rank metric, we develop an explicit rank constraint regularizer instead of the popular nuclear norm to address the rank minimization NP-hard problem.

The rest sections of this paper are organized as follows. In Section II, we present a brief discussion of the related work. Then we propose our novel robust transfer metric learning in Section III. Experimental evaluations are reported in Section IV, which is followed by the conclusion in Section V.

II. Related Work

In this part, we mainly discuss the related work in two lines, and highlight the difference of our method by comparing with existing ones.

A. Metric Learning

Metric learning [2], [20] has been popular in the area of pattern recognition and image analysis in the past decades. Most metric learning algorithms manage to learn a positive semi-definite distance matrix to boost the learning problem. Along this line, Xing *et al.* [21] developed a discriminative metric by minimizing the distances between similar pairs while keeping those of dissimilar pairs to a lower bound. Ding et al. developed a low-rank metric by exploiting the sample structure and discriminative information for face recognition [3]. However, conventional algorithms are limited by the underlying assumption that training data and test data are lying in the similar distribution. However, their appealing performance highly relies on many well-labeled training data, but they cannot deal with the challenge where there are no or limited labeled data in the target domain.

In the literature, there are several cross-domain metric learning algorithms by borrowing the knowledge from the well-labeled source domain to facilitate the unlabeled target learning [16], [17], [19]. Following this, Zha et al. developed a robust target metric by adopting a log-determinant divergence to measure the difference of source metrics and target metric [16]. Wang *et al.* [19] designed a cross-domain metric approach to borrow the source knowledge for the target domain through an information-theoretic setting. However, current cross-domain metric learning algorithms ignore the conditional distribution disparity across two domains, meanwhile they only consider to couple source and target in the sample space.

B. Transfer Learning

Transfer learning is a powerful technique to deal with the domain mismatch in many real-world scenarios [10]. The key problem of transfer learning turns to be adapting either feature space or classifiers, or both of them to bridge the distribution gap across source and target domains [14], [22]–[25].

Feature/classifier adaptation could transfer the well-labeled source knowledge to alleviate the unlabeled target learning. Over the past decades, a variety of transfer learning algorithms [13], [24]–[30], have been proposed and achieved promising performance. In this paper, we aim to seek a robust

cross-domain metric to transform source and target data to a latent shared space. We not only align source and target in sample space, but also couple them in feature space. In this way, our method could uncover more shared knowledge across two domains.

Cross-domain data matching [13] is one special case of transfer learning, which generally aims to match two different modalities of one subject to mitigate the modality divergence. Along this line, Lin et al. unified the similarity measure and feature representation learning via deep convolutional neural networks [13]. Differently, the setting of data matching problem is different from that of traditional transfer learning. In general, data matching problem would have multiple pairs of data from two modalities for training, while testing on new pairs. It is very similar to multi-view learning [25], [31]. However, in this work, we mainly concentrate on the traditional transfer learning problem where we have a well-labeled source and unlabeled target domain for training, aiming to predict the labels of the target data. Our cross-domain metric learning tends to transfer the knowledge through two directions. First of all, our method aims to minimize both marginal and conditional distribution differences across two domains simultaneously. Secondly, we adopt two strategies to make it more robust to noisy data in feature space: one is low-rank constraint and the other is marginalized desnoising scheme. Furthermore, we develop an explicit rank minimization regularizer to better recover a low-rank metric.

III. The Proposed Algorithm

In this section, we first briefly discuss the motivation to design our robust transfer metric, then provide the detail of our proposed metric through domain-wise/class-wise adaptation and two robust strategies as well. Finally, an efficient solution is developed to solve the metric learning problem.

A. Distance Metric Learning Revisit

Consider $X = \{x_1, x_2, \cdots, x_n\}$ is a set of data points, where n is data size and each $x_i \in \mathbb{R}^d$ is a sample vector with d-dimensional feature. Specifically, the set of equivalent constraints is presented by $\mathcal{S} = \{(x_i, x_j)|x_i \text{ and } x_j \text{ are from the same class}\}$, and the set of inequivalent constraints is denoted by $\mathcal{D} = \{(x_i, x_j)|x_i \text{ and } x_j \text{ are from the different classes}\}$.

Assume the distance metric is $\mathcal{M} \in \mathbb{R}^{d\times d}$, and the Mahalanobis distance of two points x_i and x_j can be defined as:

$$d_{\mathcal{M}}(x_i, x_j) = (x_i - x_j)^\top \mathcal{M}(x_i - x_j), \tag{1}$$

in which $\mathcal{M}$ is positive semi-definite ($\mathcal{M} \in \mathbb{S}_+^d$). Generally, $\mathcal{M} \in \mathbb{R}^{d\times d}$ can be decomposed into $\mathcal{M} = PP^\top$, where $P \in \mathbb{R}^{d\times r}$ and $r \leq d$ is the rank of metric $\mathcal{M}$. In this way, we can rewrite $d_{\mathcal{M}}(x_i, x_j)$ as $\|P^\top(x_i - x_j)\|_2$, which builds the connection between means metric learning and subspace learning. The general metric learning algorithm is formalized as

$$\begin{aligned} &\min_{\mathcal{M}\in\mathbb{S}_+^d} \sum_{(x_i,x_j)\in\mathcal{S}} d_{\mathcal{M}}(x_i, x_j), \\ &\text{s.t.} \sum_{(x_i,x_j)\in\mathcal{D}} d_{\mathcal{M}}(x_i, x_j) \geq 1. \end{aligned} \tag{2}$$

With $\mathcal{M}$ learned in Eq. (2), the distance between the dissimilar pairs is enlarged while the distance between the similar pairs is pulled close to each other. To this end, a discriminative metric would be learned to boost the learning problem when we have enough label information. However, traditional metric learning would be invalid when there are insufficient or no labeled data.

B. Motivation

Traditional metric learning approaches have limitations, when there are sparse or no labeled data in the target domain [19], [27]. However, we could find well-labeled source domains to help the unlabeled target learning. But this practice would involve a challenging problem in conventional metric learning. Therefore, it is essential to bridge the gap between two domains with large distribution difference during learning the metric.

Previous cross-domain metric learning algorithms only adopted to minimize the marginal distribution difference through K-L divergence [19] or Maximum Mean Discrepancy [15], [27]. Therefore, they would ignore the conditional distribution of two domains, since it is also very important to leverage the conditional divergence across two domains. Besides, current cross-domain metric algorithms only consider the sample space alignment to couple source and target while ignoring the feature space alignment. Hence, they are not very effective in cross-domain data learning.

To this end, we propose domain-wise and class-wise adaptation to seek an effective cross-domain metric to mitigate the domain shift in sample space. Furthermore, current cross-domain metric learning cannot well handle the noisy data in real world. Intuitively, we propose two strategies to build a robust metric. Specifically, low-rank constraint is introduced to uncover more common feature structure across two domains. Moreover, marginalized denoising scheme is intuitively incorporated into our robust metric learning during data reconstruction. These two strategies not only generate a robust denoising metric, but also uncover more shared feature structures across two domains, which is complementary to our domain-wise/class-wise adaption in sample space.

C. Robust Transfer Metric Learning

Given a source domain X_s with n_s labeled samples from c classes: $X_s = \{(x_{s,1}, y_{s,1}), \cdots, (x_{s,n_s}, y_{s,n_s})\}$, where $x_{s,i} \in \mathbb{R}^d$ and $y_{s,i} \in [1, c]$ is its class label. Assume X_t is a target domain with n_t unlabeled samples: $X_t = \{x_{t,1}, \cdots, x_{t,n_t}\}$, where $x_{t,i} \in \mathbb{R}^d$.

1) Domain-Wise Adaptation: When X_s and X_t are drawn from different domains, it is very essential to minimize the marginal distribution across two domains by learning an effective cross-domain metric. This issue is of particular importance and gains its popularity in transfer learning. Lots of recent research activities adopt the criterion Maximum Mean Discrepancy (MMD) to measure the distribution across two domains [15], [27], that is, the means of two domains tend to be pulled close together.

To that end, we develop *domain-wise adaptation* to guide the metric learning by leveraging mean discrepancy of two domains as follows:

$$\min_{\mathcal{M}\in\mathbb{S}_+^d} (\mu_s - \mu_t)^\top \mathcal{M}(\mu_s - \mu_t) = \mathrm{tr}(\Phi\mathcal{M}), \tag{3}$$

where μ_s is the mean of X_s ($\mu_s = \frac{1}{n_s}\sum_{i=1}^{n_s} x_{s,i}$) and μ_t is the mean of X_t ($\mu_t = \frac{1}{n_t}\sum_{i=1}^{n_t} x_{t,i}$). $\mathrm{tr}(\cdot)$ is the trace operator of the matrix and $\Phi = (\mu_s - \mu_t)(\mu_s - \mu_t)^\top$. With the mean difference of two domains minimized, the marginal divergence of two domains is reduced in an unsupervised fashion. However, it cannot ensure the conditional distributions across domains also to be close, if we only mitigate the difference in the marginal distributions. Unfortunately, it is hard to measure the conditional distribution if we have no labeled data from the target domain.

2) Class-Wise Adaptation: To address this problem, we manage to utilize the pseudo labels of the target data [32] by adopting some basic classifiers. This strategy aims to uncover the underlying structure of two domains by transferring the local information. So far, we can measure conditional distributions across two domains with the following *class-wise adaptation* term:

$$\min_{\mathcal{M}\in\mathbb{S}_+^d} \sum_{i=1}^{c} (\mu_s^i - \mu_t^i)^\top \mathcal{M}(\mu_s^i - \mu_t^i) = \sum_{i=1}^{c} \mathrm{tr}(\Phi^i\mathcal{M}), \tag{4}$$

where μ_s^i is the mean of i-th class in source domain and μ_t^i is the mean of i-th class in pseudo labeled target domain, which is updated iteratively during the cross-domain metric optimization. $\Phi^i = (\mu_s^i - \mu_t^i)(\mu_s^i - \mu_t^i)^\top$.

Specifically, we first obtain pseudo labels for target domain via the classifier trained on source domain using the original features, and utilize these labels for discriminant metric learning. Then in the new metric space, we can still propagate the pseudo labels to the target domain again, but with more accuracy since we already include both domain-wise and class-wise matching. Furthermore, the newly learned labels will provide a better description for the underlying conditional distributions. We discuss class-wise adaption with iterative refinement in the later solution section.

In reality, we may have many inexact pseudo labels for target due to the large differences of two domains. However, we can still mitigate the conditional distributions through the class-wise MMD [32]. The key of class-wise adaptation to transfer the intrinsic structure from source to target is to further mitigate the divergence across two domains. To this end, we have learned a cross-domain metric in leveraging marginal/conditional distribution divergences across two domain at the same time.

To sum up, we can achieve the domain-wise and class-wise adaptation terms into the same formula as:

$$\min_{\mathcal{M}\in\mathbb{S}_+^d} \sum_{i=0}^{c} (\mu_s^i - \mu_t^i)^\top \mathcal{M}(\mu_s^i - \mu_t^i) = \sum_{i=0}^{c} \mathrm{tr}(\Phi^i\mathcal{M}), \tag{5}$$

where we define $\Phi^0 = \Phi$ for simplicity. Through EM-like refinement of RTML, a more accurate labeling could be obtained for the unlabeled target data. Therefore, if we adopt this labeling knowledge to refine our metric iteratively, then we can alternatively achieve better labeling quality.

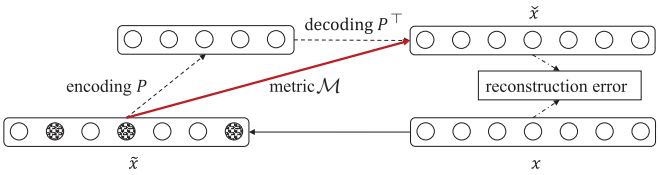


Fig. 2. Illustration of robust denoising metric, where each point x is randomly corrupted to $\tilde{x}$, then reconstructed with metric $\mathcal{M}$ to $\breve{x} = \mathcal{M}\tilde{x}$, where the metric merges encoding and decoding into one step. The loss function aims to minimize the reconstructed $\breve{x}$ and original x.

3) Robust Metric Learning Via Marginalized Denoising Fashion and Rank Constraint: It would be desirable if the metric (projection) would keep as more information available in X_s and X_t as possible. To implement this, we could formalize metric data reconstruction in a PCA-like way to preserve energy of two domains as:

$$\begin{aligned}\Omega_d &= \|X_s - PP^\top X_s\|_{\mathrm{F}}^2 + \|X_t - PP^\top X_t\|_{\mathrm{F}}^2 \\ &= \|X_s - \mathcal{M}X_s\|_{\mathrm{F}}^2 + \|X_t - \mathcal{M}X_t\|_{\mathrm{F}}^2 \\ &= \|X - \mathcal{M}X\|_{\mathrm{F}}^2,\end{aligned} \tag{6}$$

where $X = [X_s, X_t] \in \mathbb{R}^{d\times n}$, $(n = n_s + n_t)$ and $\|\cdot\|_{\mathrm{F}}$ is matrix Frobenius norm. Motivated by denoising data reconstruction, e.g., marginalized Denoising Auto-Encoder (mDAE) [11] and Denoising Auto-Encoder (DAE) [33], we tend to make our learned metric $\mathcal{M}$ with denoising property. In this way, we intuitively introduce the corrupted data in the reconstruction (6). Therefore, we could rewrite Eq.(6) into the denoising data reconstruction fashion:

$$\Omega_d = \|\bar{X} - \mathcal{M}\tilde{X}\|_{\mathrm{F}}^2, \tag{7}$$

where $\bar{X}$ is m-times repeated version of X, and $\tilde{X}$ is the corrupted version of $\bar{X}$ with different ratios of random corruption [11]. We could observe that our metric is constrained in a data reconstruction way by transforming noisy data to clean data and digging out the noise. In this way, our metric is much more robust in dealing with real data.

Remark: Our marginalized denoising metric is different from mDAE [11], which aims to learn denoising transformation (similar to P decomposed from our metric $\mathcal{M} = PP^\top$). Moreover, our denoising metric is very similar to DAE [33] (Figure 2), since $P^\top\tilde{X}$ can be treated as a linear version of the hidden layer (*encoding*), while $\mathcal{M}\tilde{X} = PP^\top\tilde{X}$ can be treated as the reconstructed output from corrupted input $\tilde{X}$ (*decoding*). To this end, we finally achieve a DAE-like robust metric for cross-domain learning. Therefore, our metric could not only have denoising property but also reduce the dimensionality of the original data to save computational cost, by comparing with [11].

Previous work assumes metric $\mathcal{M}$ to be low-rank [34], [35], so that the metric $\mathcal{M}$ can uncover the feature space structure of the data, and reduce the storage need as well as efficient distance computation. For cross-domain data, the structured metric could also transfer the well-learned knowledge from source to target to boost its performance.

Finally, we formulate our objective function for robust transfer metric learning as follows:

$$\min_{\mathcal{M}\in\mathbb{S}_+^d} \sum_{i=0}^{c} \mathrm{tr}(\Phi^i\mathcal{M}) + \alpha\|\bar{X} - \mathcal{M}\tilde{X}\|_{\mathrm{F}}^2 + \lambda \mathrm{rank}(\mathcal{M}), \quad (8)$$

where α and λ are two balanced parameters. $\mathrm{rank}(\mathcal{M})$ is the rank operator of the matrix $\mathcal{M}$. With the above objective function (8), we can learn a discriminative and robust metric for knowledge transfer, which not only uncovers the shared information in sample space, but also learns more robust structure shared by two domains in feature space.

4) Explicit Rank Control Regularization: Rank minimization is an NP-hard problem, which cannot be easily addressed. In the literature, there are a lot of strategies to find a surrogate to solve the rank minimization problem [36], [37]. One popular strategy is to adopt a surrogate nuclear norm $\|\mathcal{M}\|_*$ to replace $\mathrm{rank}(\mathcal{M})$ [36], [38]. Specifically, nuclear norm calculates the sum of all singular values of $\mathcal{M}$. Furthermore, our framework enforces $\mathcal{M}$ to be positive semi-definite (PSD) $\mathbb{S}_+^d$, and therefore, we could obtain $\|\mathcal{M}\|_* = \mathrm{tr}(\mathcal{M})$.

Note that trace operator $\mathrm{tr}(\mathcal{M})$ is defined to be the sum of the elements on the main diagonal. However, if the non-zero elements on the main diagonal of matrix $\mathcal{M}$ change, $\mathrm{tr}(\mathcal{M})$ will change as well, but the rank of $\mathcal{M}$ may keep constant. Therefore, nuclear norm (trace operator) is not a good surrogate for rank minimization problem.

To that end, we develop a novel term that minimizes the sum of the $d-r$ smallest eigenvalues of $\mathcal{M} \in \mathbb{S}_+^d$ into Eq. (8) so that the novel term would reach its minimum when the rank of $\mathcal{M}$ is less or equal to a pursued rank r:

$$\min_{\mathcal{M}\in\mathbb{S}_+^d} \sum_{i=0}^{c} \mathrm{tr}(\Phi^i\mathcal{M}) + \alpha\|\bar{X} - \mathcal{M}\tilde{X}\|_{\mathrm{F}}^2 + \lambda \sum_{i=r+1}^{d} \big(\sigma_i(\mathcal{M})\big)^2, \quad (9)$$

where $\sigma_i(\mathcal{M})$ is the i-th eigenvalue of $\mathcal{M}$. Such a minimization controls the rank of $\mathcal{M}$ in an explicit manner, since the rank of the PSD matrix $\mathcal{M} \in \mathbb{S}_+^d$ equals the number of its non-zero eigenvalues.

Specifically, we can notice that $\sum_{i=r+1}^{d} \big(\sigma_i(\mathcal{M})\big)^2 = \mathrm{tr}(\Gamma^\top\mathcal{M}\mathcal{M}^\top\Gamma)$, where Γ are the singular vectors which correspond to the $(d-r)$-smallest singular values of $\mathcal{M}\mathcal{M}^\top$. To sum up, we have the final objective function as:

$$\min_{\mathcal{M}\in\mathbb{S}_+^d,\Gamma} \sum_{i=0}^{c} \mathrm{tr}(\Phi^i\mathcal{M}) + \alpha\|\bar{X} - \mathcal{M}\tilde{X}\|_{\mathrm{F}}^2 + \lambda \mathrm{tr}(\Gamma^\top\mathcal{M}\mathcal{M}^\top\Gamma), \quad (10)$$

a) Discussion: To sum up, we build a robust cross-domain metric for effective target learning through knowledge transfer in two directions. To mitigate the marginal and conditional disparity across two domains in sample space, we propose domain-wise and class-wise adaptation terms to better transfer the knowledge from source to target. With the iterative refinement strategy, we could involve more accurate labels of the target for valid knowledge transfer. For feature space, our metric learning could be integrated into a DAE-style data reconstruction so that our metric would be more robust by manually introducing the noise into the training data [33]. Furthermore, we develop an explicit rank minimization regularizer to better achieve a low-rank metric so that we could reduce the redundant features across two domains. In a word, the learned metric is more effective and robust in dealing with unlabeled target learning.

D. Optimization

With positive semi-definite constraint, it cannot directly optimize the metric $\mathcal{M}$. First of all, we define $h(\mathcal{M},\Gamma) = \sum_{i=0}^{c} \mathrm{tr}(\Phi^i\mathcal{M}) + \alpha\|\bar{X} - \mathcal{M}\tilde{X}\|_{\mathrm{F}}^2 + \lambda \mathrm{tr}(\Gamma^\top\mathcal{M}\mathcal{M}^\top\Gamma)$.

Specifically, we adopt a linear approximation to $h(\mathcal{M},\Gamma)$ to solve the problem [35]. In this way, $\mathcal{M}$ and Γ tend to be optimized in an iterative manner with leaving-one-out scheme. Also for each iteration, we need to refine the pseudo labels of the target data to optimize the class-wise adaption term by involving more accurate labels of target data. That is, we optimize one by fixing the other iteratively. Define $\mathcal{M}_t$ and Γ_t are the optimization at time t, then the $\mathcal{M}_{t+1}$ and Γ_{t+1} are updated at the $(t+1)^{\text{th}}$ iteration:

1) Update $\mathcal{M}$:

$$\begin{aligned}
\mathcal{M}_{t+1} &= \arg\min_{\mathcal{M}\in\mathbb{S}_+^d} h(\mathcal{M},\Gamma_t) \\
&= \arg\min_{\mathcal{M}\in\mathbb{S}_+^d} \frac{1}{2\eta}\|\mathcal{M} - \mathcal{M}_t\|_F^2 + h(\mathcal{M}_t,\Gamma_t) \\
&\quad + \langle \nabla_{\mathcal{M}} h(\mathcal{M},\Gamma_t)|_{\mathcal{M}=\mathcal{M}_t}, \mathcal{M} - \mathcal{M}_t\rangle \\
&= \arg\min_{\mathcal{M}\in\mathbb{S}_+^d} \frac{1}{2\eta}\|\mathcal{M} - (\mathcal{M}_t - \eta\mathcal{H}_t)\|_{\mathrm{F}}^2 \\
&= \mathcal{P}_{\mathbb{S}_+^d}(\mathcal{M}_t - \eta\mathcal{H}_t), \quad (11)
\end{aligned}$$

where $\mathcal{H}_t = \nabla_{\mathcal{M}} h(\mathcal{M},\Gamma_t)|_{\mathcal{M}=\mathcal{M}_t} = \sum_{i=0}^{c}\Phi^i + 2\alpha(\bar{X}\tilde{X}^\top - \mathcal{M}_t\tilde{X}\tilde{X}^\top) + 2\lambda\Gamma_t\Gamma_t^\top\mathcal{M}_t$ and $\eta > 0$ is the step size.

In fact, we would like the repeated number m to be ∞, hence, the marginal denoising metric $\mathcal{M}_{t+1}$ could be effectively obtained from infinite copies of noisy data. Actually, when m tends to be large enough, $\mathcal{U} = \tilde{X}\tilde{X}^\top$ and $\mathcal{V} = \tilde{X}\tilde{X}^\top$ can converge to their expectations under the weak law of large numbers [11]. Therefore, we can update $\mathcal{H}_t$ by calculating the exceptions of $\mathcal{U}/\mathcal{V}$ as:

$$\begin{aligned}
\mathcal{H}_t &= \mathbb{E}[\sum_{i=0}^{c}\Phi^i + 2\alpha(\mathcal{U} - \mathcal{M}_t\mathcal{V}) + 2\lambda\Gamma_t\Gamma_t^\top\mathcal{M}_t] \\
&= \mathbb{E}[\sum_{i=0}^{c}\Phi^i + 2\lambda\Gamma_t\Gamma_t^\top\mathcal{M}_t] + \mathbb{E}[2\alpha(\mathcal{U} - \mathcal{M}_t\mathcal{V})] \\
&= \sum_{i=0}^{c}\Phi^i + 2\lambda\Gamma_t\Gamma_t^\top\mathcal{M}_t + 2\alpha\big(\mathbb{E}[\mathcal{U}] - \mathcal{M}_t\mathbb{E}[\mathcal{V}]\big), \quad (12)
\end{aligned}$$

where $\mathbb{E}[\mathcal{U}] = \Big(\sum_i x_i x_i^\top \otimes \Lambda_u\Big)$ and $\mathbb{E}[\mathcal{V}] = \Big(\sum_i x_i x_i^\top \otimes \Lambda_v\Big)$. Λ_u (Λ_v) is $d \times d$ matrix encoding joint survival ratios of two single features from original data x_i and corrupted data $\tilde{x}_i$ (both from corrupted data $\tilde{x}_i$), and $\otimes$ is the element-wise multiplication [11].

Furthermore, the introduced operator $\mathcal{P}_{\mathbb{S}_+^d}(\cdot)$ denotes the projection operation to $\mathbb{S}_+^d$. Specifically, for a symmetric matrix $\mathcal{K} \in \mathbb{R}^{d\times d}$, $\mathcal{P}_{\mathbb{S}_+^d}(\mathcal{K}) = \sum_{i=1}^{d}[\gamma_i]_+ k_i k_i^\top$, in which $(k_i\gamma_i)_{i=1}^d$ are its eigenvector-eigenvalue pairs.

2) Update Γ*:* When $\mathcal{M}_{t+1}$ is optimized, we could update Γ_{t+1} with the eigenvectors corresponding to the $(d-r)$-smallest singular values of $\mathcal{M}_{t+1}\mathcal{M}_{t+1}^\top$. To compute Γ_{t+1}, we have to perform singular value decomposition (SVD) of matrix $\mathcal{M}_{t+1}\mathcal{M}_{t+1}^\top$. Suppose singular value decomposition of $\mathcal{M}_{t+1}\mathcal{M}_{t+1}^\top = U_\mathcal{M}\Sigma_\mathcal{M}U_\mathcal{M}^\top$, where $U_\mathcal{M}$ is the eigenvector matrix and $\Sigma_\mathcal{M}$ is the diagonal matrix in ascending order. Denote $U_\mathcal{M} = [U_\mathcal{M}^1, U_\mathcal{M}^2]$, in which $U_\mathcal{M}^1 \in \mathbb{R}^{d\times(d-r)}$, and $U_\mathcal{M}^2 \in \mathbb{R}^{d\times r}$, then we can directly achieve $\Gamma_{t+1} = U_\mathcal{M}^1$.

According to the Eq. (12), to update $\mathcal{M}$, we do not need the exact value of Γ, instead, we only need to calculate $\Gamma\Gamma^\top$. It is easy to know that $U_\mathcal{M}U_\mathcal{M}^\top = U_\mathcal{M}^1U_\mathcal{M}^{1\top} + U_\mathcal{M}^2U_\mathcal{M}^{2\top} = \mathrm{I}_d$. We have $\Gamma\Gamma^\top = \mathrm{I}_d - U_\mathcal{M}^2U_\mathcal{M}^{2\top}$. We know that $\mathcal{M}\mathcal{M}^\top$ is a low-rank matrix, so r is a small value ($r \ll d$). Since the previous procedure would cost $\mathcal{O}((d-r)^2d) \approx \mathcal{O}(d^3)$ for matrix multiplication of $\Gamma_{t+1}\Gamma_{t+1}^\top$. While our optimized procedure would take $\mathcal{O}(r^2d) \approx \mathcal{O}(d)$ for matrix multiplication of $U_\mathcal{M}^2U_\mathcal{M}^{2\top}$. Therefore, it is more efficient than directly calculating $U_\mathcal{M}^1$.

3) Update $\Phi^i, 1 \le i \le c$*:* After we have optimized the low-rank metric $\mathcal{M}$ for each iteration, we can decompose $\mathcal{M} = PP^\top$ to achieve the low-dimensional projection. Assume the rank of $\mathcal{M}$ is $r \ll d$, especially for high-dimensional data, therefore $P \in \mathbb{R}^{d\times r}$ can project the original cross-domain data to an r-dimensional space for dimension reduction. To achieve P, we run the eigen-decomposition of $\mathcal{M} = U_p\Sigma_pU_p^\top$, where $\Sigma_p \in \mathbb{R}^{r\times r}$ only preserves the non-negative eigenvalues. Therefore, $P = U_p\Sigma_p^{\frac{1}{2}}$. When P is learned, we could apply it to extract low-dimensional features from two domains, then we adopt the nearest neighbor classifier (NNC) to predict the labels of the target data. To this end, the class-wise adaptation term would be updated with new predicted labels of the target data.

We would optimize all the variables iteratively until the metric converges, i.e., $\|\mathcal{M}_{t+1} - \mathcal{M}_t\|_\infty < \epsilon$. The detail solution to the proposed Robust Transfer Metric Learning (RTML) approach is summarized in **Algorithm 1**. Specifically, we empirically set the parameters $\eta = 0.01$ and $\epsilon = 10^{-4}$, while tuning λ and α in the experimental part.

E. Complexity Analysis

In this part, we would discuss the complexity of our algorithm. There are four major consuming parts:

- Matrix multiplication in Step 1;
- SVD-projection in Step 2;
- SVD decomposition in Step 3;
- Refine class-wise adaptation in Step 4

Next, we would discuss the detail complexity of these four steps. For $\mathcal{H}_t$ (Step 1), there are several matrix multiplications, e.g., l multiplications, so it would cost $\mathcal{O}(ld^3)$. Step 2 takes $\mathcal{O}(d^3)$ when projecting $\mathcal{M}$ onto $\mathbb{S}_+^d$ through SVD-based projection. For Step 3, the SVD decomposition would take $\mathcal{O}(d^3)$ and the calculation of Γ needs around $\mathcal{O}(d)$. While Step 4, it usually takes $\mathcal{O}(cn^2)$ to update the class-wise adaptation matrix. To sum up, the whole complexity of RTML is bounded with $\mathcal{O}\Big(\mathrm{T}\big((l+2)d^3 + cn^2 + d\big)\Big)$, where T is the optimization iteration number.

Algorithm 1 Robust Transfer Metric Learning

Input: labeled source data $\{X_s, y_s\}$, unlabeled target data X_t, $\alpha, \lambda, \eta = 0.01, \epsilon = 10^{-4}, \mathrm{T} = 100$ and $\Phi^i, 0 \le i \le c$.
Initialize: $\mathbb{E}(\mathcal{U}), \mathbb{E}(\mathcal{V})$;
while not converged **or** $t \le \mathrm{T}$ **do**
1. $\mathcal{H}_t = \sum_{i=0}^{c}\Phi^i + 2\alpha(\bar{X}\tilde{X}^\top - \mathcal{M}_t\tilde{X}\tilde{X}^\top) + 2\lambda\Gamma_t\Gamma_t^\top\mathcal{M}_t$;
2. $\mathcal{M}_{t+1} = \mathcal{P}_{\mathbb{S}_+^d}(\mathcal{M}_t - \eta\mathcal{H}_t)$;
3. Update Γ_{t+1};
4. Update $\Phi^i, 1 \le i \le c$;
5. Check the convergence condition $\|\mathcal{M}_{t+1} - \mathcal{M}_t\|_\infty < \epsilon$.
6. $t = t + 1$.

end while
output: $\mathcal{M}$.

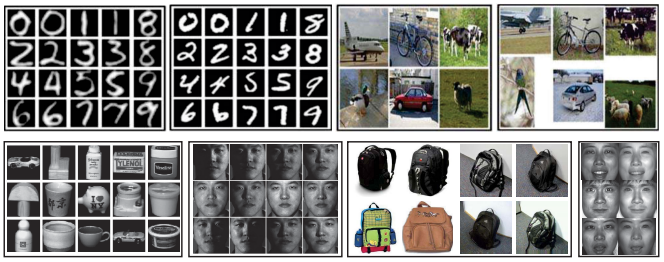


Fig. 3. Image Samples from left to right are USPS digit, MNIST digit, MSRC object and VOC 2007 object, COIL20 object, CMU-PIE face, Office+Caltech256 and BUAA VIS-NIR face, respectively.

IV. Experiments

In this section, we first present the benchmarks and experimental setting. Then comparison results are presented followed by some property analysis.

A. Datasets & Experimental Setting

Office+Caltech256, CMU-PIE, USPS, MNIST, COIL20, MSRC, VOC2007 and BUAA VIS-NIR (see Figure 3 and Table I) are 9 image benchmark datasets widely adopted.

MRSC+VOC contains two subsets: (1) MSRC dataset[1] has more than 4000 samples from 18 categories; (2) VOC2007 dataset[2] includes over 5000 samples annotated with 20 concepts. Six shared categories are selected to build MSRC+VOC (1269 samples in MSRC and 1530 samples in VOC2007). Dense SIFT features are used with 128 dimensions.

USPS+MNIST[3] shares ten common digit categories from two subsets: USPS and MNIST. In this experiment, we use 1800 samples from USPS and 2000 samples from MNIST. All the images are rescaled to 16×16, and the raw pixel values are adopted.

[1] http://research.microsoft.com/en-us/projects/objectclassrecognition.
[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007.
[3] http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html

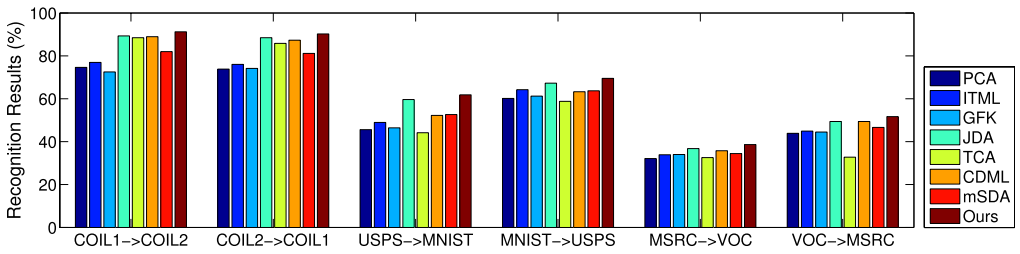


Fig. 4. Recognition results of 8 algorithms on different databases: COIL20, USPS+MNIST digit, MRSC+VOC object.

TABLE I
STATISTICS OF THE 9 BENCHMARK DATASETS

Dataset	Type	#Example	#Features	#Class
COIL20	Object	1,440	1,024	20
USPS	Digit	1,800	256	10
MNIST	Digit	2,000	256	10
CMU-PIE	Face	11,554	1024	68
MSRC	Photo	1,269	240	6
VOC2007	Photo	1,530	240	6
Office	Object	1,410	800/4,096	10
Caltech-256	Object	1,123	800/4,096	10
BUAA VIS-NIR	Face	1,350	900	150

COIL20 object dataset[4] includes 20 objects with 1440 images. Each object has 72 images with each image of size 32×32. In the experiments, we follow the setting of [41] to split the dataset into 2 subsets: COIL1 and COIL2, in which COIL1 includes samples of $[0°, 85°] \cup [180°, 265°]$ and COIL2 includes samples of $[90°, 175°] \cup [270°, 355°]$. We select one subset as one domain, then build two transfer learning task.

CMU PIE Face database consists of 68 subjects with each under 21 various illumination conditions. We adopt five pose subsets: C05, C07, C09, C27, C29, which provides a rich basis for domain adaptation, that is, we can choose one pose as the source and any rest one as the target. Therefore, we obtain $5 \times 4 = 20$ different source/target combinations. Finally, we combine all five poses together to form a single dataset for large-scale transfer learning experiment. We crop all images to 32×32 and only adopt the pixel values as the input.

Office+Caltech256[5] contains 10 shared categories from Office dataset and Caltech-256. Specifically, Office dataset included three subsets, i.e., Amazon, Webcam, and DSLR. Caltech-256 is a standard object dataset with 256 categories over 30000 samples. In this dataset, we adopt two kinds of features, i.e., 800-dim SURF and 4096-dim DeCAF[6] features. We select two out of four as one cross-domain task. Finally, we conduct 4×3 different groups of domain adaptation experiments. We adopt the sampling protocol [43] that randomly samples the source domain with 20 labeled examples per category for Amazon (A) and Caltech (C) and 8 labeled examples per category for Webcam (W) and DSLR (D). While the target domain is totally unlabeled.

BUAA VIS-NIR face database is adopted for heterogeneous knowledge transfer with two modalities. Specifically, we randomly choose 75 subjects and their corresponding one modality images as one domain, and use the remaining 75 subjects with the images from the other modality. Since there is no overlap across two domains, therefore, we select one image per subject in the target domain as the reference. Five selection on different reference images are conducted and we reported the average performance.

Note that the arrow "→" is the direction from "source" to "target". For example, "Webcam → DSLR" means Webcam is the labeled source domain and DSLR is the unlabeled target.

B. Comparison Methods & Implementation Details

We mainly compare with 7 state-of-the-art methods to show the effectiveness of our algorithm as follows: Principle Component Analysis (PCA) [39], Information-theoretic Metric Learning (ITML) [40], Geodesic Flow Kernel (GFK) [23], Joint Domain Adaptation (JDA) [41], Transfer Component Analysis (TCA) [42], Cross-Domain Metric Learning (CDML) [19] and Marginalized Denoising Auto-encoder (mSDA) [11].

The first two are the traditional metric learning algorithms, in which we train the metric on labeled source data wile reuse for target learning. The last five are the transfer learning algorithms and CDML is the transfer metric learning one. For transfer learning algorithms, we train on labeled source and unlabeled target to transfer knowledge during the model training.

In transfer learning, it is hard to tune the optimal parameters through cross validation. Therefore, we empirically search the optimal parameter, and report each method's best results. Furthermore, for the pseudo labels initialization, we adopt the labeled source data to predict the unlabeled target using the nearest neighbor classifier (NNC) with the original features. Note that different initializations, e.g., random initialization, would influence on the final performances.

C. Experimental Results

In this section, we present the comparison results on different datasets including digit, object and face images, to show the effectiveness of our proposed algorithm.

We first experiment on cross-domain object databases, e.g., digit, object images. For COIL20, MNIST+USPS and MSRC+VOC, each has two subsets, so we select one as the source domain while the other as the target domain, then we switch them. In all, we have two cases for each database, and the comparison results of 8 algorithms are shown in Figure 4. For Office+Caltech256, we strictly follow [23] to repeat 20 times and calculate the average performance as

[4]http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[5]http://www-scf.usc.edu/ boqinggo/domainadaptation.html

TABLE II
RECOGNITION RATE OF 8 ALGORITHMS ON CMU-PIE CROSS POSE FACE DATASET

Config\Methods	PCA [39]	ITML [40]	GFK [23]	JDA [41]	TCA [42]	CDML [19]	mSDA [11]	Ours
C05→C07	25.43	27.75	26.15	58.81	40.76	53.22	28.35	60.12
C05→C09	26.87	27.33	27.27	54.23	41.79	53.12	26.91	55.21
C05→C27	30.98	31.60	31.15	84.50	59.63	80.12	30.39	85.19
C05→C29	17.21	19.00	17.59	49.75	29.35	48.23	21.76	52.98
C07→C05	25.13	26.05	25.24	57.62	41.81	52.39	28.27	58.13
C07→C09	47.43	48.71	47.37	62.93	51.47	54.23	44.19	63.92
C07→C27	53.98	55.54	54.25	75.82	64.73	68.36	55.39	76.16
C07→C29	27.12	29.53	27.08	39.89	33.70	37.34	28.08	40.38
C09→C05	21.67	22.99	28.69	50.96	34.69	43.54	24.83	53.12
C09→C07	42.87	44.20	43.16	57.95	47.70	54.87	42.59	58.67
C09→C27	45.97	47.34	46.41	68.45	56.23	62.76	50.25	69.81
C09→C29	26.43	28.25	26.78	39.95	33.15	38.21	27.83	42.13
C27→C05	34.12	35.83	34.24	80.58	55.64	75.12	32.89	81.12
C27→C07	62.09	64.46	62.92	82.63	67.83	80.53	63.10	83.92
C27→C09	72.67	74.39	73.35	87.25	75.86	83.72	74.70	89.51
C27→C29	37.43	39.46	37.38	54.66	40.26	52.78	34.81	56.26
C29→C05	21.08	21.40	20.35	23.17	26.98	27.34	25.85	29.11
C29→C07	24.61	25.72	24.62	31.74	29.90	30.82	26.33	33.28
C29→C09	28.19	29.72	28.49	38.17	29.90	36.34	28.63	39.85
C29→C27	31.05	31.93	31.33	45.99	33.64	40.61	32.98	47.13
Average	35.13	36.56	35.69	57.25	44.75	53.69	36.41	58.80

TABLE III
AVERAGE RECOGNITION RATE (%)± STANDARD DEVIATION OF 8 ALGORITHMS ON OFFICE+CALTECH-256 (SURF FEATURES & DECAF$_6$), WHERE A = AMAZON, D = DSLR, C = CALTECH-256 AND W = WEBCAM

Config\Methods	PCA [39]	ITML [40]	GFK [23]	JDA [41]	TCA [42]	CDML [19]	mSDA [11]	Ours
C→W (SURF)	33.9±0.5	43.3±0.7	40.7±0.3	41.7±0.4	30.5±0.5	35.6±0.8	33.6±0.6	43.5±0.5
C→D (SURF)	35.2±0.8	42.4±1.1	38.9±0.9	45.2±0.8	35.7±0.5	42.5±0.4	38.3±0.4	45.5±0.6
C→A (SURF)	36.9±0.7	46.2±0.6	41.1±0.6	44.8±0.7	39.0±0.6	47.7±0.6	38.8±0.8	49.7±0.4
A→C (SURF)	35.6±0.5	35.3±0.8	40.3±0.4	39.4±0.5	39.1±0.7	40.7±0.6	31.3±0.4	42.7±0.5
A→W (SURF)	34.4±0.7	38.6±0.6	39.0±0.9	37.8±0.3	35.3±0.8	37.3±0.7	35.5±0.5	43.4±0.9
A→D (SURF)	34.9±0.6	37.6±0.7	36.2±0.7	39.5±0.7	34.4±0.6	35.3±0.5	29.7±0.7	43.3±0.6
W→C (SURF)	27.3±0.7	32.3±0.4	30.7±0.1	31.2±0.4	29.9±0.3	31.6±0.4	30.4±0.5	36.9±0.5
W→A (SURF)	31.3±0.6	33.4±0.5	29.8±0.6	32.8±0.6	28.8±0.6	32.4±0.5	32.1±0.8	37.5±0.7
W→D (SURF)	70.7±0.5	80.3±0.8	80.9±0.4	89.2±0.9	86.0±1.0	77.9±0.9	56.6±0.4	91.7±1.1
D→C (SURF)	31.7±0.6	32.4±0.6	31.5±0.5	31.6±0.9	32.1±0.5	32.2±0.5	31.4±0.4	37.0±0.5
D→A (SURF)	33.6±0.3	38.0±0.9	33.2±0.6	33.1±0.7	31.4±0.8	29.4±0.8	33.6±0.8	36.3±0.3
D→W (SURF)	83.2±0.4	83.6±0.6	79.4±0.6	89.5±0.8	86.4±0.5	79.4±0.6	68.6±0.7	90.5±0.7
Average (SURF)	40.7	45.3	43.5	46.3	42.4	43.5	38.4	49.8
C→W (DeCAF$_6$)	66.1±1.6	68.8±1.7	71.6±1.4	83.7±1.1	79.3±1.3	77.6±1.8	67.2±1.6	83.8±1.2
C→D (DeCAF$_6$)	74.5±2.2	78.4±2.5	82.4±2.9	86.6±2.4	76.2±1.9	84.5±3.3	76.1±3.5	88.7±2.3
C→A (DeCAF$_6$)	85.6±1.9	86.6±1.8	87.6±1.5	89.7±1.7	87.7±1.9	88.7±2.3	82.0±3.8	90.2±1.9
A→C (DeCAF$_6$)	70.3±1.4	72.6±1.7	74.8±1.3	82.2±1.7	73.0±1.9	78.5±1.7	72.8±1.8	83.1±1.6
A→W (DeCAF$_6$)	57.2±2.5	65.2±1.3	73.1±2.8	78.6±1.4	62.3±1.8	75.9±2.1	64.6±4.2	79.5±2.6
A→D (DeCAF$_6$)	64.9±3.7	73.8±1.7	82.6±2.1	80.2±2.5	68.4±2.7	81.4±2.6	72.6±3.5	83.8±1.7
W→C (DeCAF$_6$)	60.3±1.8	66.5±1.4	72.6±1.7	80.5±1.6	67.4±2.2	78.0±1.2	69.1±1.9	82.9±1.8
W→A (DeCAF$_6$)	62.5±1.3	72.5±1.2	82.6±1.3	88.1±1.5	71.2±1.4	86.3±1.6	71.4±1.7	90.8±1.6
W→D (DeCAF$_6$)	98.7±1.2	98.8±0.5	98.8±0.9	100±0.6	98.2±0.8	99.4±0.4	99.5±0.6	100±0.5
D→C (DeCAF$_6$)	52.0±0.9	62.7±0.4	73.3±0.8	80.1±0.3	62.1±0.3	79.2±0.8	74.7±0.4	81.6±0.5
D→A (DeCAF$_6$)	62.7±0.7	74.1±0.2	85.4±0.7	89.4±0.4	65.9±0.6	88.4±0.5	78.8±0.5	90.6±0.5
D→W (DeCAF$_6$)	89.1±0.7	90.2±0.8	91.3±0.4	98.9±0.5	89.2±0.4	95.1±0.5	97.9±0.4	98.6±0.3
Average (DeCAF$_6$)	70.4	75.9	81.4	86.5	75.1	84.5	77.3	87.8

well as the variation. The comparison results are presented in Table III. We further compare our algorithm on cross-pose face database, i.e., CMU-PIE (Table II). There are five poses, each could be one domain, so we could build 20 cases by randomly selecting 2 for one.

PCA and ITML are two traditional metric learning algorithms, which train the model on source data while evaluate on target data. Specifically, PCA is totally unsupervised while ITML needs side information (supervised knowledge). We could observe that such algorithms cannot perform well since the source and target have domain shift. In this way, the metric trained on source cannot be well reused in the target domain.

GFK designs a kernel metric to minimize the divergence of source and target. GFK outperforms other comparisons in some cases, e.g., Office+Caltech database. The reason may be kernel metric could well mitigate the distribution gap in these type of data. TCA proposes a unified framework to learn a projection by matching feature representation. Differently, JDA builds a unified dimensionality reduction algorithm to

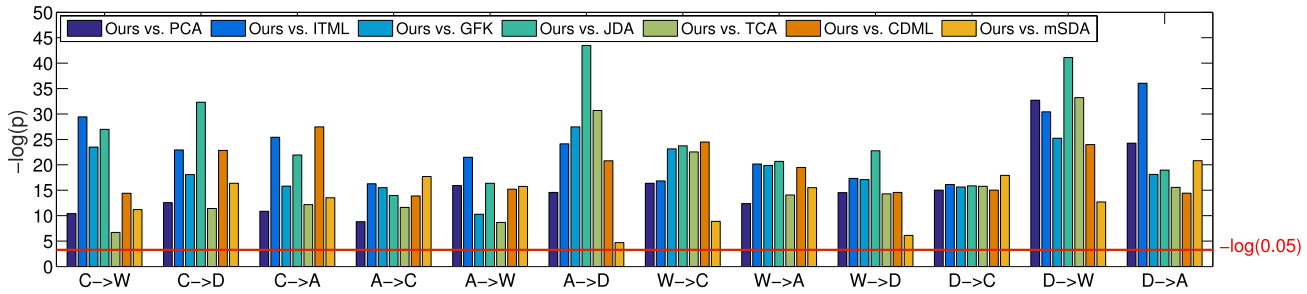


Fig. 5. p-value of t-test between our method and others on Office+Caltech256 (SURF features). We do pre-processing using $-\log(p)$ so that the large value shown in the figure means the more significance of our algorithm compared with others.

TABLE IV

AVERAGE RECOGNITION RATE (%)± STANDARD DEVIATION OF 8 ALGORITHMS ON BUAA NIR-VIS FACE DATABASE

Config\Methods	PCA [39]	ITML [40]	GFK [23]	JDA [41]	TCA [42]	CDML [19]	mSDA [11]	Ours
VIS→NIR	30.0±0.5	37.1±0.7	38.1±0.3	48.3±0.4	36.8±0.5	45.3±0.8	38.4±0.6	49.5±0.5
NIR→VIS	59.1±0.8	61.2±1.1	59.2±0.9	61.8±0.8	59.4±0.8	60.1±0.4	59.6±0.4	62.1±0.9
Average	44.5	49.1	48.6	55.0	48.1	52.7	49.0	55.8

TABLE V

RECOGNITION RATE OF 4 ALGORITHMS ON DIFFERENT EVALUATION CASES

Config\Methods	RTML$_l$	RTML$_c$	RTML$_d$	Ours
COIL1→COIL2	90.87	80.87	89.23	91.23
USPS→MNIST	60.16	48.36	59.96	61.82
MSRC→VOC	37.87	36.23	37.03	38.63
C05→C09	54.97	36.54	54.26	55.21
C→W (SURF)	42.7±0.4	40.7±0.3	43.2±0.5	43.5±0.5
W→A (SURF)	36.6±0.4	30.8±0.5	36.2±0.5	37.5±0.7
A→C (DeCAF$_6$)	82.7±1.5	74.4±1.4	82.0±1.3	83.1±1.6
W→C (DeCAF$_6$)	82.4±1.8	73.3±1.5	81.8±1.7	82.9±1.8

mitigate the marginal/conditional distributions. These algorithms all adopt the shallow structure. From the comparison results, we notice that JDA can achieve better results in most cases by comparing with other shallow structure transfer learning. The key reason is that JDA not only involves the source labels into model training, but also iteratively optimizes the target labels for the class-wise adaptation term. Similarly, our proposed algorithm also involves pseudo labels of target to iteratively optimize the class-wise adaption term. However, JDA only considers the sample direction alignment across source and target, while ignores the feature direction matching. Our method intuitively incorporates the marginalized denoising and low-rank strategies into our metric learning, which would make our metric more robust to noise cases [11]. Moreover, such strategies could further uncover more shared features for two domains to boost the knowledge transfer.

CDML is a cross-domain metric which adopts the K-L divergence to measure the similarity of two domains in order to mitigate the disparity across two domains. However, CDML only considers the marginal distribution difference of two domains. This K-L divergence strategy cannot well align source and target to transfer more effective knowledge. From the results, we could notice that our proposed algorithm could consistently outperform CDML, since our algorithm not only involves the class-wise adaption term, but also works in a marginalized denoising fashion. In this way, our metric would be more robust and effective deal with real-world images. Although mSDA adopted denoising strategy, it only combined source and target together to seek new representation. Hence, the domain shift could not be well mitigated.

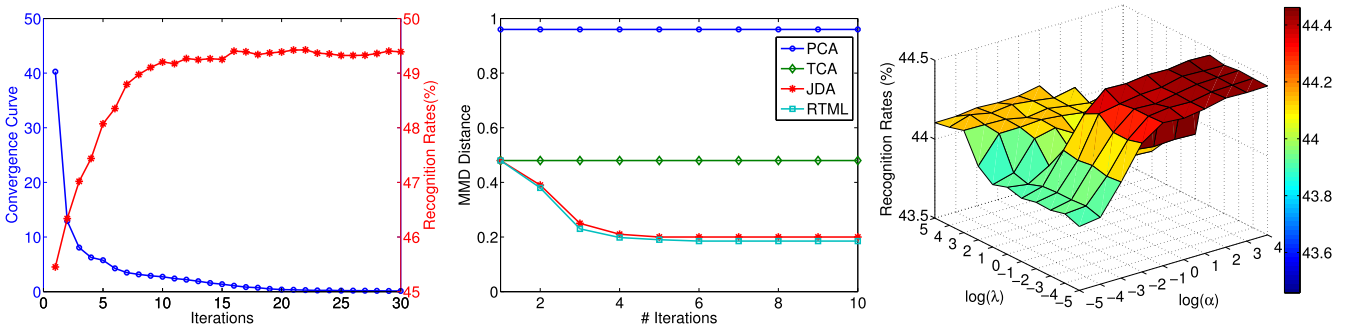
From the results of Table III, we could observe that all the algorithms achieve better results using DeCAF$_6$ cases comparing with using SURF. This verifies the deep features have better discriminative information. The domain shift across source and target has been mitigated to some extents using the same deep structure. In some challenging cases, e.g., $C \rightarrow A$ and $A \rightarrow C$, deep features show much better performance by comparing with SURF features. But in some easy cases, $D \rightarrow W$ and $W \rightarrow D$, deep features do not show much superiority over SURF features. In both features, our method can outperform others, which shows the effectiveness of the proposed method in knowledge transfer.

Furthermore, to demonstrate the statistical significance of our approach, we also performed a significance test (t-test) for the results shown in Figure 5. We used a significance level of 0.05. That is to say, when p-value is less than 0.05, the performance difference of two methods is statistically significant. Figure 5 lists the p-values of our method by comparing with others. Since we do $-\log(p)$ processing, the comparison shows that our method outperforms others significantly if the values are greater than $-\log(0.05)$.

D. Parameter Analysis

In this section, we testify some properties of the proposed RTML, i.e., component evaluation, convergence analysis, MMD distance, influence of two parameters and training time.

First of all, to understand our model deeply, we evaluate several variants, i.e., (1) RTML$_l$ by removing the low-rank constraint, (2) RTML$_c$ by removing the class-wise adaptation, (3) RTML$_d$ by removing the marginalized denoising part. We evaluate on different cases following the previous setting and the results are shown in Table V. We could observe that RTML$_c$ performs worse than the other two variants and



(a) (b) (c)

Fig. 6. (a) Recognition curve (red '*') and convergence curve (Blue 'o') of our linear algorithm on Office+Caltech256 $C \rightarrow W$. (b) MMD distance of four algorithms on $A \rightarrow W$ on Office+Caltech256 database. (c) Parameters analysis results of λ and α on Office+Caltech256 database for the case $C \rightarrow D$. The values of x-axis and y-axis are used log() to rescale the size.

TABLE VI

TRAINING TIME (*second*) OF RTML AND ALL THE BASELINE METHODS

Methods	PCA [39]	ITML [40]	GFK [23]	JDA [41]	TCA [42]	CDML [19]	mSDA [11]	Ours
C05→C07	2.63	4.32	4.58	46.32	3.82	12.42	6.79	69.23
C→A (DeCAF$_6$)	25.75	52.94	59.79	134.69	30.76	87.32	50.43	149.65
USPS→MNIST	1.26	1.65	1.73	38.66	1.46	5.21	3.43	40.22
COIL1→COIL2	1.72	3.27	3.43	41.09	2.43	8.28	5.02	51.27

RTML$_l$ works better than other two variants. However, all the three variants cannot achieve better results than RTML.

Secondly, we experiment on convergence and recognition performance of our approach in different iterations. Specifically, we set $\alpha = 10$ and $\lambda = 10^{-1}$. The case $C \rightarrow A$ on Office+Caltech256 database is used for evaluation and the results are presented in Figure 6(a). We could notice that our algorithm converges very well, meanwhile the recognition performance goes up quickly and reaches to a stable position.

Thirdly, we evaluate the MMD distance of PCA, TCA, and JDA on dataset $A \rightarrow W$. It is worth to note that better generalization performance could be obtained if the distribution distance is smaller. Figure 6(b) lists MMD distances of four approaches, where we could see the MMD distance of RTML is the smallest. The reason is that RTML is able to mitigate the marginal/conditional distributions more effectively by building a most discriminative and robust metric. Through EM-like refinement, we could observe that RTML reduces the MMD distance iteratively.

Moreover, we evaluate on setting MSRC→VOC. The influence of parameters shows the recognition performance at different values in Figure 6(c), where we could observe that larger α and smaller λ generate better performance. α shows more important impact by comparing with λ, that is, our marginalized denoising term would play an important role in learning an effective cross-domain metric. Besides, low-rank constraint could also avoid the scale issue, as it keeps the metric in certain spaces with limited magnitude. We observe the performance would degrade when $\lambda = 0$ from Table V.

Finally, we evaluate the efficiency of the algorithms on four cases from different cross-domain databases, and show the results in Table VI. For JDA and RTML, we run 10 iterations so that they are more time-consuming. We observe that our algorithm and JDA cost more time than other comparisons. The key reason is we both adopt the iterative refinement to optimize the problem. In each iteration, we need to calculate the target's labels.

V. CONCLUSION

In this paper, we developed a Robust Transfer Metric Learning (RTML) framework through low-rank constraint and marginalized denoising scheme. The cross-domain metric guided by domain-wise and class-wise adaption terms was incorporated to align source and target domains, by minimizing the marginal and conditional distributions across two domains. Furthermore, two strategies were developed to make our learned cross-domain metric more robust to noisy data, one is the low-rank constraint on metric and the other is the marginalized denoising scheme during data reconstruction. Specifically, low-rank constraint tended to preserve the feature structure while marginalized denoising strategy was designed to better handle corrupted data in real world. Experimental results verified the superiority of the designed cross-domain metric.

REFERENCES

[1] R. Jin, S. Wang, and Y. Zhou, "Regularized distance metric learning: Theory and algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 862–870.

[2] A. Bellet, A. Habrard, and M. Sebban. (2013). "A survey on metric learning for feature vectors and structured data." [Online]. Available: https://arxiv.org/abs/1306.6709

[3] Z. Ding, S. Suh, J.-J. Han, C. Choi, and Y. Fu, "Discriminative low-rank metric learning for face recognition," in *Proc. 11th IEEE Int. Conf. Autom. Face Gesture Recognit.*, vol. 1. May 2015, pp. 1–6.

[4] Q. Wang, W. Zuo, L. Zhang, and P. Li, "Shrinkage expansion adaptive metric learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 456–471.

[5] S. Wang and R. Jin, "An information geometry approach for distance metric learning," in *Proc. Artif. Intell. Statist. Conf.*, 2009, pp. 591–598.

[6] L. Yang, R. Jin, R. Sukthankar, and Y. Liu, "An efficient algorithm for local distance metric learning," in *Proc. Assoc. Adv. Artif. Intell.*, vol. 2, 2006, pp. 543–548.

[7] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[8] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 309–316.

[9] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou, "Neighborhood repulsed metric learning for kinship verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 331–345, Feb. 2014.

[10] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[11] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 767–774.

[12] D. Mandal and S. Biswas, "Generalized coupled dictionary learning approach with applications to cross-modal matching," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3826–3837, Aug. 2016.

[13] L. Lin, G. Wang, W. Zuo, F. Xiangchu, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, [Online]. Avaliable: http://ieeexplore.ieee.org/document/7469374/

[14] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative transfer subspace learning via low-rank and sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 850–863, Feb. 2016.

[15] B. Geng, D. Tao, and C. Xu, "DAML: Domain adaptation metric learning," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2980–2989, Oct. 2011.

[16] Z.-J. Zha, T. Mei, M. Wang, Z. Wang, and X.-S. Hua, "Robust distance metric learning with auxiliary knowledge," in *Proc. 21st Int. Jont Conf. Artif. Intell.*, 2009, pp. 1327–1332.

[17] Y. Zhang and D.-Y. Yeung, "Transfer metric learning with semi-supervised extension," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, 2012, Art. no. 54.

[18] Y. Luo, T. Liu, D. Tao, and C. Xu, "Decomposition-based transfer distance metric learning for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3789–3801, Sep. 2014.

[19] H. Wang, W. Wang, C. Zhang, and F. Xu, "Cross-domain metric learning based on information theory," in *Proc. Assoc. Adv. Artif. Intell.*, 2014, pp. 2099–2105.

[20] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, 2013.

[21] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Proc. Neural Inf. Process. Syst.*, 2002, pp. 505–512.

[22] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2960–2967.

[23] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073.

[24] S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa, "Generalized domain-adaptive dictionaries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 361–368.

[25] Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 110–119.

[26] Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4322–4334, Nov. 2015.

[27] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 325–333.

[28] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.

[29] L. Duan, D. Xu, and I. W. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 504–518, Mar. 2012.

[30] J. Ni, Q. Qiu, and R. Chellappa, "Subspace interpolation via dictionary learning for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 692–699.

[31] Y. Yan, E. Ricci, S. Subramanian, G. Liu, and N. Sebe, "Multitask linear discriminant analysis for view invariant action recognition," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5599–5611, Dec. 2014.

[32] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1076–1089, May 2014.

[33] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.

[34] G. Zhong, K. Huang, and C.-L. Liu, "Low rank metric learning with manifold regularization," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 1266–1271.

[35] W. Liu, C. Mu, R. Ji, S. Ma, J. R. Smith, and S.-F. Chang, "Low-rank similarity metric learning in high dimensions," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2792–2799.

[36] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[37] M. T. Law, N. Thome, and M. Cord, "Fantope regularization in metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1051–1058.

[38] J. Li, Y. Kong, H. Zhao, J. Yang, and Y. Fu, "Learning fast low-rank projection for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4803–4814, Oct. 2016.

[39] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[40] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, pp. 209–216, 2007.

[41] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2200–2207.

[42] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[43] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 213–226.

Zhengming Ding (S'14) received the B.Eng. degree in information security and the M.Eng. degree in computer software and theory from the University of Electronic Science and Technology of China, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, USA. His research interests include machine learning and computer vision. Specifically, he devotes himself to develop scalable algorithms for challenging problems in transfer learning scenario. He is an AAAI Student Member. He was a recipient of the Student Travel Grant of ACM MM 14, ICDM 14, AAAI 16, and IJCAI 16, and the best paper award (SPIE). He received the National Institute of Justice Fellowship. He has served as a Reviewer of the IEEE journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

Yun Fu (S'07–M'08–SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign. Since 2012, he has been an Interdisciplinary Faculty Member with the College of Engineering and the College of Computer and Information Science, Northeastern University. He has extensive publications in leading journals, books/book chapters, and international conferences/workshops. His research interests are machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems. He serves as an Associate Editor, Chairs, a PC Member, and a Reviewer of many top journals and international conferences/ workshops. He received seven Prestigious Young Investigator Awards from the NAE, ONR, ARO, IEEE, INNS, UIUC, and Grainger Foundation, seven Best Paper Awards from the IEEE, IAPR, SPIE, and SIAM, and three major Industrial Research Awards from Google, Samsung, and Adobe. He is a fellow of IAPR, a Lifetime Senior Member of ACM and SPIE, Lifetime Member of AAAI, OSA, and the Institute of Mathematical Statistics, a member of Global Young Academy, INNS, and the Beckman Graduate Fellow from 2007 to 2008. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEANING SYSTEMS.