Learning Discriminatively Reconstructed Source Data for Object Recognition with Few Examples

Pai-Heng Hsiao, Feng-Ju Chang, and Yen-Yu Lin, *Member, IEEE*

***Abstract*—We aim at improving object recognition with few training data in the *target* domain by leveraging abundant auxiliary data in the *source* domain. The major issue obstructing knowledge transfer from source to target is the limited correlation between the two domains. Transferring irrelevant information from the source domain usually leads to performance degradation in the target domain. To address this issue, we propose a transfer learning framework with the two key components, *discriminative source data reconstruction* and *dual-domain boosting*. The former correlates the two domains via reconstructing source data by target data in a discriminative manner. The latter discovers and delivers only knowledge shared by the target data and the reconstructed source data. Hence, it facilitates recognition in the target. The promising experimental results on three benchmarks of object recognition demonstrate the effectiveness of our approach.**

***Index Terms*—Object recognition, domain adaptation, transfer learning, low-rank reconstruction, boosting, late fusion.**

I. Introduction

OBJECT recognition has always been a fundamental yet critical problem in the fields of image processing and computer vision, since it is essential to a broad spectrum of applications, such as image retrieval, semantic image segmentation, and scene understanding. However, object recognition has become more and more challenging in nowadays applications owing to increasing numbers of objects to be identified, large intra-class variations, as well as expensive manner efforts for training data acquisition and labeling.

Conventional works employ various powerful hand-made feature descriptors or machine learning methodologies to conquer these difficulties. One famous combination is *multiple kernel learning* (MKL), e.g., [1]–[7], which can fuse multiple, complementary descriptors to enhance object recognition. Multiple kernel learning requires a large number of labeled data. However, training data are often insufficient due to the expensive manual efforts for training data collection and labeling.

To reduce manual efforts, *transfer learning* [8] has been largely developed and demonstrated its effectiveness for handling classification with few data in various applications, such as object recognition [9]–[14], image segmentation [15], object tracking [16], image denoising [17], and so on. The idea behind transfer learning is to deliver the beneficial knowledge in the *source* to improve the learning task in the *target*. Recent research advances, e.g., [10]–[13], design various ways to transfer distinct source information, and are usually accompanied by mechanisms for avoiding *negative transfer*, which is caused by passing meaningless or harmful information from source to target.

Existing transfer learning algorithms [10]–[13] have presented the efficacy of solving the problem of few labeled data in object recognition, but most of them still suffer from some limitations. First, they usually assume strong correlation between the source and target domains. For instance, some methods [18], [19] hypothesize that the conditional probabilities of labels given the samples in the two domains are similar. Second, they handle merely the knowledge transfer from either a single class or multiple classes in the source to a single class in the target. However, discriminative information over classes in the target domains is preferred and even more important. These restrictions reduce their performance and applicability for object recognition, and might lead to the negative transfer.

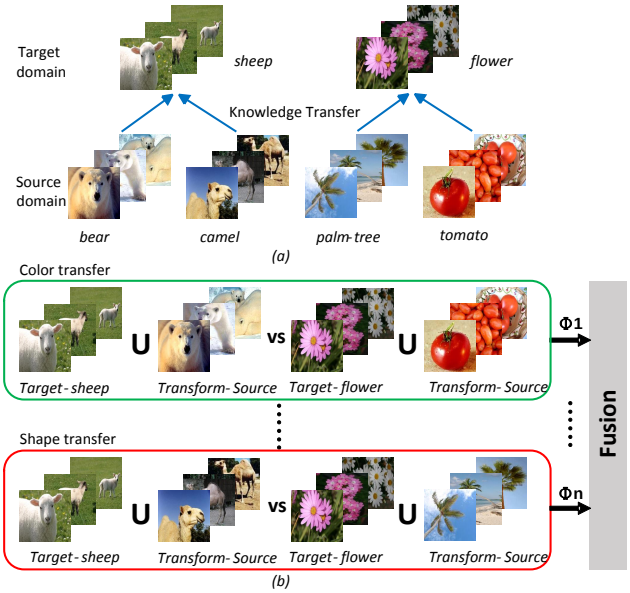


Fig. 1. Overview of our approach. (a) In the case where two target classes and four source classes are given, data of each source class is reconstructed by data of either target class, indicated by the arrows. (b) For each adopted feature representation such as color in the green box, each target class is augmented with the reconstructed source data. Dual-domain boosting is used to learn the one-vs-one classifier. After learning the classifier for each feature representation, late fusion is employed for prediction.

This work was supported by Ministry of Science and Technology (MOST) and Institute for Information Industry (III) under Grant MOST 103-2221-E-001-026-MY2, Grant MOST 104-2628-E-001-001-MY2, and Grant III 105-EC-17-A-24-0691.

P.-H. Hsiao, and Y.-Y. Lin are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan. (E-mail: ylsd214@gmail.com, yylin@citi.sinica.edu.tw).

F.-J. Chang is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, U.S.A. (E-mail: fengjuch@usc.edu).

In this paper, we propose a general transfer learning approach that delivers relevant information from multiple source classes to multiple target classes for object recognition, and solves the issues mentioned above. The main contributions of this work are summarized as follows. First, we do not assume that there exists strong correlation between source and target domains. In our approach, data of the two domains can be from either the same or different databases, and their respective object categories may partially overlap or don't overlap at all. To enhance the correlation between the two domains, our approach reconstructs source data by target data via a discriminative, low-rank formulation. Only the reconstructed source data are used for helping recognition in the target. In this way, the inter-domain variations are reduced to correlate the two domains. By discriminative reconstruction, we mean that data of a source class is enforced to be reconstructed mostly by data of a single target class. It enhances the discriminative power in transfer. Second, after correlating two data domains, we transfer knowledge from the reconstructed source data to the target domain. The developed *dual-domain boosting* can explore common knowledge in the two domains, and leverages the extra knowledge to improve recognition in the target. Specifically, it jointly learns two boosted classifiers, one for each domain, in which the relatedness between domains are modeled by the *shared weak learners* and the differences are reflected by their *respective ensemble coefficients*. The proposed boosting algorithm has its theoretic support. While the shared weak learners are selected to best minimize the total exponential loss of both source and target domains, the ensemble coefficients are derived to optimally minimize the respective loss. Third, multi-class recognition in the target is carried out by using *one-vs-one* strategy in this work. We augment each one-vs-one classifier with a *two-vs-rest* classifier, i.e., the two involved classes versus the rest, and show that the performance can be further improved via diminishing the influence of irrelevant votes in the one-vs-one scheme.

Fig. 1 gives the overview of our method. In Fig. 1(a), two target classes and four source classes are considered. Data of each source class are reconstructed by data of either target class, as indicated by the arrows. In the green box of Fig. 1(b), each target class is augmented with the reconstructed source data before performing one-vs-one classification via dual-domain boosting. The procedure is repeated for each adopted feature, such as the color-based and shape-based features. A late fusion is employed for final prediction.

We assess the proposed method on three benchmark datasets, including Caltech256 [20], SUN09 [21], and MSRC [22]. The promising performance of *within-database transfer*, i.e., the source and target classes coming from the same dataset, and *cross-database transfer*, i.e., the source and target classes coming from different datasets, demonstrates the effectiveness of our approach, even when only few target training data (ten or fewer examples per class) are available.

II. Related Work

In this section, we review a couple of research topics that are relevant to the development of our approach, including object recognition and transfer learning.

A. Object Recognition

Object recognition is typically formulated as a multi-class classification problem. Vast research effort has centered on minimizing intra-class variations while maximizing inter-class variations by designing more powerful features to distinguish objects of a class from another. Since there is no single universal descriptor in general that can well represent all objects to be recognized, multiple kernel learning algorithms (MKL), e.g., [3], [5], are employed to investigate the optimal fashion for integrating various features.

The major limitation of MKL is that they need a large number of labeled training data to build stable classifiers. Unluckily, it is usually hard to label or compile ample data, particularly for new object categories. Therefore, exploring the knowledge in correlated categories from different domains to learn a new object category with limited supervision has become an active research issue in object recognition, also called transfer learning.

B. Transfer Learning

The spirit of transfer learning is to identify the domain-specific and domain-sharing knowledge in the source domain and transfer the latter to improve the task of the target domain with little supervision. The recent survey of transfer learning algorithms can be found in [8], [23]. Based on [8], these algorithms can be divided into four categories according to what kind of source information is transferred:

1) Transfer by data instances: Methods of this class, e.g., [10], [13], [24], aim to re-weight or re-sample source data instances by measuring the correlation of samples between the source and target domains via various techniques, such as boosting [10], [13] or data distribution matching [24]. The source training instances that are similar to the target ones are given larger weights for increasing their impacts on target model learning.

Knowledge transfer in this way relies on high correlation between instances of source and target domains. Our approach belongs to methods of this category, but can be distinguished from others with the difference that the correlation in this approach is *reconstructed* and hence enhanced by using a discriminative, low-rank formulation in which source instances are reconstructed by target instances. The reconstructed source instances are leveraged by dual-domain boosting for transfer learning. Thus, we can extend the feasibility of transfer learning without worrying about whether there are enough similar and correlated instances in source and target domains in the beginning.

2) Transfer by model parameters: The model transfer for object recognition is pioneered by Fei-Fei et al. [11]. They suggest to learn an object category with a small training set via updating the general knowledge, which is extracted from previously learned categories and represented as a prior probability function in the space of model parameters. Some of research efforts for model transfer are based on the support vector machine (SVM) or its variants [12], [25]–[27]. The primary assumption of these methods is that the closely related source and target domains will have similar parameters or

priors of models. Hence, the target classifier can be obtained by adjusting the pre-learned classifiers in the source domain via the few training examples. Luo et al. [28] instead propose distance metric transfer by assuming that the target distance metric can be optimized by searching the space spanned by the eigenvectors of source distance metrics.

These works assume that both source and target domains share similar model parameters. The models adapt the existing source classifiers by fitting target data. However, such schemes may have limited applicability especially when source and target domains are dissimilar.

3) Transfer by relational knowledge: This kind of transfer learning is suitable for multi-label classification e.g., visual concept detection [29]. The major hypothesis is that data drawn from each domain is not independent and identically distritbuted (i.i.d.) [8]. Instead, the visual concepts in an image are often positively or negatively correlated. For instance, fishes must be in water so that if we are given an image with a fish, then the probability of seeing water is also very high. Jiang et al. [29] exploit a semantic graph to capture the correlation between concepts, and update it gradually based on the concept relatedness in the target domain. Qi et al. [30] introduce a cross-category transfer learning algorithm exploring concept correlation between categories. One requirement in this type of transfer is that these methods depend on a large number of human-driven multi-label data.

4) Transfer by feature representation: Approaches of this category attempt to find out more discernible feature representations to recognize objects with the aid of auxiliary source data. Various ways have been proposed to carry out this type of transfer, such as learning intermediate feature representations [31]–[37], e.g., attributes and binary codes, exploring invariant domains or subspaces [18], [19], [38]–[43], deriving a number of classifiers from auxiliary data [44]–[46], and learning a cross-domain dictionary pair by which the sparse codes of source and target data are correlated [47], [48].

Attribute-based transfer learning algorithms aim to bridge the low-level image features and high-level concepts. However, expensive human efforts on labeling attributes [32]–[34], [36] are required. Research efforts [18], [19], [38]–[40] are centered on exploring invariant domains or subspaces that align the feature distributions between source and target domains. Nevertheless, these methods often address only cross-domain object recognition with the same categories in the source and target, and assume the conditional distributions of labels in the two domains are similar. Another notable branch for this type of transfer is to build a number of classifiers on source data. New feature representations of target data are yielded by applying these classifiers to the data, e.g., *classemes* [44]. However, exploring good features among such a rich set of intermediate features to address the target task still relies on a large number of training data in practice.

III. The Proposed Approach

We formulate object recognition as a multi-class classification problem in the target domain, and adopt *one-vs-one* scheme to carry out it. It leads to $\frac{C_T(C_T-1)}{2} \cdot M$ binary classification problems, where C_T is the number of classes in the target and M is the number of the adopted feature representations. For each binary problem, we focus on learning a binary predictor by using data in the *two* corresponding categories in the target and *all* categories in the source. Specifically, this task is carried out by the three components of our transfer learning approach. First, we correlate the two domains by discriminative reconstruction in the manner that data of each source class is reconstructed mainly by data of either target class. Each of the two target classes is then augmented with the respectively reconstructed and correlated source data. Second, we use the developed boosting algorithm to explore and deliver knowledge from source to target. Further, a *two-vs-rest* classifier is learned to exclude the influence of irrelevant voters in the one-vs-one strategy for multi-class classification. Third, a *late fusion* mechanism is used to fuse the predictions for data in multiple feature representations.

The rest of this section is organized as follows. We give the notations and the problem definition in subsection III-A. To correlate the target and source domains by discriminative reconstruction, our formulation and its optimization are introduced in subsections III-B and III-C, respectively. The dual-domain boosting is specified in subsection III-D. Finally, late fusion for working with multiple feature representations is described in subsection III-E.

A. Notations and Problem Definition

We aim at improving image category recognition in the target domain with few training examples by borrowing information from an abundant set of data in the source domain. We are given a training set in the target domain, $D_T = \{(\mathbf{x}_n^T \in \mathcal{X}, y_n^T \in \mathcal{Y}_T)\}$ of C_T classes, as well as a large dataset in the source domain, $D_S = \{(\mathbf{x}_n^S \in \mathcal{X}, y_n^S \in \mathcal{Y}_S\}$ of C_S classes. In this work, the categories of the target and source datasets, $\mathcal{Y}_T$ and $\mathcal{Y}_S$, can partially overlap, or have completely different categories. Even for data in a common category, their feature distributions in the two domains may be different. To deal with complex recognition tasks, we use M descriptors to better characterize each sample in both two domains, i.e., $\mathbf{x}_n = \{\mathbf{x}_{n,m} \in \mathcal{X}_m\}_{m=1}^M$. Our goal is to derive a better classifier that gives low generalization error in the target domain by leveraging not only information extracted from D_T but also knowledge transferred from D_S.

B. Domain Correlation via Discriminative Reconstruction

In the one-vs-one scheme, two object classes in the target and all classes in the source under one particular feature representation are considered. The goal in this step is two-fold. First, we need to correlate data in the two domains, since no assumption about the correlation of the two domains is made in advance. Second, we aim to enrich the training data in the target by borrowing data in the source, so we divide each of source object classes into either side of the two target classes based on their data similarities, and carry out enrichment.

We develop a formulation of *discriminative reconstruction* to simultaneously accomplish the two aforementioned tasks, i.e., domain correlation and data augmentation. Inspired by

the good performance reported in the pioneering work by Jhuo et al. [40], reconstruction-based domain adaptation is adopted in our approach. In this way, data of each source class are reconstructed by data of the two target classes via minimizing the reconstruction error. The correlation of the two domains is then performed by borrowing information only from the reconstructed part of the source data. We go beyond the formulation [40] by using discriminative reconstruction. It means that data of this source class are enforced to be reconstructed by data of either target class, instead of both classes. More discriminative information can be borrowed for learning one-vs-one classifiers in this way. The procedure is repeated for each source class. All source classes are then *binarized*, and each source class is then assigned to the target class reconstructing it.

Our approach to discriminative reconstruction is introduced in the following. Without loss of generality, we assume that the two target object categories are classes 1 and 2 and the single source category is class c. Their object instances are under feature representation m. We aggregate the data of two target categories, and denote them by $\{X_\ell^T \in \mathbb{R}^{d\times N_\ell}\}_{\ell=1}^2$, where d is the dimension of representation m, and N_ℓ is the number of data in target class ℓ. The data of source class c are similarly aggregated $X^S \in \mathbb{R}^{d\times N}$, where N is the number of data in source class c. The discriminative reconstruction is repeated for each source class and each adopted feature representation. The indices c and m are hence omitted for simplicity.

The inter-domain variations may be large, and make knowledge transfer infeasible. We assume that the unfavorable variations can be modeled by a linear transformation, and the related source data can be reconstructed by target data after transformation. The idea can be specified by

$$WX^S = X^TZ + E, \tag{1}$$

$$= [X_1^T\ X_2^T]\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + E, \tag{2}$$

where $W \in \mathbb{R}^{d\times d}$ is the transformation matrix, $Z = [\mathbf{z}_1\ \mathbf{z}_2\ \cdots\ \mathbf{z}_N] \in \mathbb{R}^{(N_1+N_2)\times N}$ is the matrix of the reconstruction coefficients, and $E = [\mathbf{e}_1\ \mathbf{e}_2\ \cdots\ \mathbf{e}_N] \in \mathbb{R}^{d\times N}$ is the matrix of the reconstruction errors. It can be checked column by column in Eq. (1) that each transformed source sample $W\mathbf{x}_i^S$ is well reconstructed by target data $X^T\mathbf{z}_i$, if residue $\mathbf{e}_i$ is minimized. In Eq. (2), X^T and Z are divided into two parts according to the target classes. These coefficients specify how the source data are reconstructed by either the whole or part of target data. Since we aim to improve the discriminative power of reconstructed source data, the reconstruction coefficients of data of each source class are restricted to be related to only one side of target data.

It is not practical to solve W, Z, and E in Eq. (1) directly, since there are infinite feasible solutions. Therefore, a low-rank based formulation is adopted and serves as the prior to solve these variables. Besides, we implement discriminative reconstruction by minimizing the minimal norm of Z_1 and Z_2 so that source data tend to be reconstructed by only data of a single target class. The task can be cast as the following constrained optimization problem:

$$\min_{W,Z,E} \quad \mathrm{rank}(Z) + \alpha\|E\|_{2,1} + \min(\|Z_1\|_F^2, \|Z_2\|_F^2) \tag{3}$$

$$s.t. \quad WX^S = X^TZ + E, \tag{4}$$

$$WW^\top = I, \tag{5}$$

where $\|E\|_{2,1} = \sum_{i=1}^N \|\mathbf{e}_i\|_2$ is the $l_{2,1}$ norm of E, α is a positive tradeoff parameter, and $\|Z_\ell\|_F^2$ is the square of Frobenius norm of Z_ℓ. Constraint $WW^\top = I$ ensures that W is a basis transformation, i.e., an orthogonal matrix. Minimizing the rank of Z helps the preservation of the intrinsic structure of X^S. The use of $l_{2,1}$ norm in the error measure enforces consistently small errors in the features, but tolerates large reconstruction errors caused by outliers, i.e., source data that cannot be precisely reconstructed. Thus, it alleviates the overfitting problem caused by the outliers. The term $\min(\|Z_1\|_F^2, \|Z_2\|_F^2)$ carries out discriminative reconstruction. It minimizes the smaller reconstruction coefficient matrix between $\|Z_1\|$ and $\|Z_2\|$ in terms of Frobenius norm. Namely, this term encourages that source data X^S are mainly reconstructed by either X_1^T or X_2^T, but not both.

Rank minimization in general is known as an NP-hard problem, and there is no efficient algorithm to solve it. Hence, we consider the convex relaxation of the optimization problem in Eq. (3), i.e.,

$$\min_{W,Z,E} \quad \|Z\|_* + \alpha\|E\|_{2,1} + \min(\|Z_1\|_F^2, \|Z_2\|_F^2) \tag{6}$$

$$s.t. \quad WX^S = X^TZ + E, \tag{7}$$

$$WW^\top = I, \tag{8}$$

where $\|Z\|_*$ is the nuclear norm of Z, i.e., the sum of its singular values. It serves as a convex approximation of $\mathrm{rank}(Z)$.

To solve the constrained optimization problem in Eq. (6), we use the *Augmented Lagrange Multiplier* (ALM) method [49], which deals with a constrained optimization problem by solving a series of unconstrained ones. First of all, we convert the optimization problem in Eq. (6) to an equivalent form

$$\min_{W,Z,Z',Z'',E} \quad \|Z'\|_* + \alpha\|E\|_{2,1} + \min(\|Z_1''\|_F^2, \|Z_2''\|_F^2) \tag{9}$$

$$s.t. \quad WX^S = X^TZ + E, \tag{10}$$

$$Z = Z', \tag{11}$$

$$Z = Z'', \tag{12}$$

where Z' and $Z'' = [Z_1''^\top\ Z_2''^\top]^\top$ are the additional auxiliary variables of Z, and they are required in the following optimization procedure. Note that Z is the coefficient matrix for source data X_S reconstruction. The orthogonality constraint of W is temporarily ignored. Nevertheless, W is *orthogonalized* afterwards as in most orthogonality preserving methods.

We solve the optimization problem in Eq. (9) by the *inexact ALM method*, which minimizes the *augmented Lagrange*

function of Eq. (9):

$$\begin{aligned}\min_{W,Z,Z',Z'',E,Y,Y',Y''} \quad & \|Z'\|_* + \alpha\|E\|_{2,1} + \min(\|Z_1''\|_F^2, \|Z_2''\|_F^2) \\ & + \langle Y, WX^S - X^TZ - E\rangle + \frac{\mu}{2}\|WX^S - X^TZ - E\|_F^2 \\ & + \langle Y', Z - Z'\rangle + \frac{\mu}{2}\|Z - Z'\|_F^2 \\ & + \langle Y'', Z - Z''\rangle + \frac{\mu}{2}\|Z - Z''\|_F^2, \end{aligned} \tag{13}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product operator, μ is a positive penalty parameter, and Y, Y' and Y'' are the Lagrange multipliers. Refer to [50] for the details of the augmented Lagrange function.

How to optimize Eq. (13) will be specified in the next subsection. Suppose that the optimization is completed currently. The source data of this class is then assigned to target class with the smaller reconstruction error, i.e., to target class $\ell^* = \arg\min_{\ell\in\{1,2\}} \|WX^S - X_\ell^T Z_\ell\|_F$. The reconstructed part of the source data of this class is denoted by $X^R = [\mathbf{x}_1^R\ \mathbf{x}_2^R\ \cdots\ \mathbf{x}_N^R] = X_\ell^T Z_\ell$, where each reconstructed source sample is assigned to class ℓ. For each source class, we repeat above optimization procedure, and assign the reconstructed data of that source class to the target class with smaller reconstruction error. It follows that all the reconstructed source data are correlated with the target data and binarized, i.e., $\{(\mathbf{x}_i^R, y_i^R \in \{1,2\})\}_{i=1}^{N_R}$, where N_R is the number of the all source data.

C. Domain Correlation Optimization

We give the details how the objective function Eq. (13) is solved in this subsection. Starting with a small value of μ, the inexact ALM method iteratively solves Eq. (13) by gradually enlarging μ. The procedure is repeated until all the constraints in Eq. (10) $\sim$ (12) are satisfied. At each iteration, the strategy of alternating optimization is adopted for solving variables $\{Z', Z'', W, E, Z\}$. Namely, we optimize one of the variables by fixing the rest, and then switch roles of the variables sequentially. Lagrange multipliers, Y, Y' and Y'', are updated accordingly. The subproblems for variables $\{Z', Z'', W, E, Z\}$ and the optimization methods are introduced as follows:

a) On optimizing Z': We fix all the optimization variables except Z', and yield the subproblem w.r.t. Z'. The *singular value shrinkage operator* $\mathcal{D}_\tau$ in [51] serves as the solver to optimize Z' with $\mathcal{D}_{\frac{1}{\mu}}(Z + \frac{Y'}{\mu})$:

$$Z' = \arg\min_{Z'} \frac{1}{\mu}\|Z'\|_* + \frac{1}{2}\|Z' - (Z + \frac{Y'}{\mu})\|_F^2. \tag{14}$$

b) On optimizing Z'': The resulting subproblem w.r.t. Z'' is given below:

$$\begin{aligned} Z'' = \arg\min_{Z''} \ & \min(\|Z_1''\|_F^2, \|Z_2''\|_F^2) \\ & + \langle Y'', Z - Z''\rangle + \frac{\mu}{2}\|Z - Z''\|_F^2. \end{aligned} \tag{15}$$

We would like to solve this subproblem by using gradient descent methods. However, the $\min$ operation applying to $\|Z_1''\|_F^2$ and $\|Z_2''\|_F^2$ makes the objective function non-differentiable. To address this issue, we adopt the *Log-Sum-Exp* (LSE) trick, which gives the differentiable surrogate of the $\min$ operation by

$$\begin{aligned} & \min(\|Z_1''\|_F^2, \|Z_2''\|_F^2) \approx \\ & \quad \frac{-1}{r}\log\{\exp(-r\|Z_1''\|_F^2) + \exp(-r\|Z_2''\|_F^2)\}, \end{aligned} \tag{16}$$

where the smoothness parameter r is a positive constant. It follows that the subproblem in Eq. (15) can be approximated by

$$\begin{aligned} Z'' = \arg\min_{Z''} \ & \frac{-1}{r}\log\left(\exp(-r\|Z_1''\|_F^2) + \exp(-r\|Z_2''\|_F^2)\right) \\ & + \langle Y'', Z - Z''\rangle + \frac{\mu}{2}\|Z - Z''\|_F^2. \end{aligned} \tag{17}$$

We then solve Eq. (17) by using gradient descent.

c) On optimizing W: For the subproblems w.r.t. W, there exists a closed-form solution derived by setting the partial derivative to zero. That is,

$$W = (X^TZ + E - \frac{Y}{\mu})(X^S)^\top(X^S(X^S)^\top)^{-1}. \tag{18}$$

Besides, we use QR-decomposition to orthogonalize the obtained W such that $WW^\top = I$ holds.

d) On optimizing E: The subproblem w.r.t. E requires the $l_{2,1}$ norm minimization. Following [52], it can be updated with the analytical solution, i.e.,

$$E = \arg\min_E \frac{\alpha}{\mu}\|E\|_{2,1} + \frac{1}{2}\|E - (WX^S - X^TZ + \frac{Y}{\mu})\|_F^2. \tag{19}$$

e) On optimizing Z: For the subproblem w.r.t. Z, we derive the closed-form solution by setting the partial derivative to zero:

$$\begin{aligned} Z = & (2I + (X^T)^\top X^T)^{-1}[(X^T)^\top(WX^S - E) \\ & + \frac{1}{\mu}((X^T)^\top Y - Y' - Y'') + (Z' + Z'')]. \end{aligned} \tag{20}$$

The optimization procedure is summarized in Algorithm 1. Parameter μ in Eq. (13) determines the relative importance of the constraints. When iteratively running Algorithm 1, μ is gradually enlarged until all the constraints are satisfied. Thus, μ should be set as a small value, and gradually enlarged until all constraints are satisfied. ρ is introduced to control the step size of iteratively enlarging μ. When the initial value of μ and the value of ρ are set as 10^{-3} and 1.2 respectively, the optimization procedure in Algorithm 1 converges with $45 \sim 62$ iterations in our experiments. The computational bottleneck of Algorithm 1 lies in step 3 and step 4. Hence, the running time can be significantly reduced by using principal component analysis (PCA) to preprocess data. In the case where X^T contains 20 target data, 10 for each target class, X^S has 50 source data, and each pre-processed data sample is of dimension 300, the running time of Algorithm 1 is about 1.48 seconds on a modern PC with an Intel Core $i7$ 3.4 GHz processor.

D. Knowledge Transfer via Dual-Domain Boosting

After discriminative reconstruction, the transformed source data are correlated with target data in the feature space, and aligned as well in the labels. Namely, the set of the binary-class target data is $\{(\mathbf{x}_i^T, y_i^T \in \{1,2\})\}_{i=1}^{N_T}$, and the set of

Algorithm 1: *Inexact ALM for Solving Problem in Eq. (13)*

Input: Binary-class target data X^T, source data X^S of a single class, parameter α.

Initialize: $E = 0; W = I; Y = Y' = Y'' = 0;$
$Z = ((X^T)^\top (X^T))^{-1}(X^T)^\top W X^S$; $\mu = 10^{-3}$.

while *not converged* **do**

1. Update Z' in Eq. (14) by using singular value thresholding.
2. Update Z'' in Eq. (17) by using gradient descent.
3. Update W with the closed-form solution in Eq. (18).
4. $W \leftarrow$ orthogonal(W).
5. Update E with the analytical solution Eq. (19).
6. Update Z with the closed-form solution in Eq. (20).
7. Update the Lagrange multipliers:
   $Y = Y + \mu(WX^S - X^T Z - E)$,
   $Y' = Y' + \mu(Z - Z')$,
   $Y'' = Y'' + \mu(Z - Z'')$.
8. Update μ by $\mu = \min(\mu\rho, 10^{10})$, where ρ = 1.2.
9. Check the convergence conditions:
   $WX^S - X^T Z - E \to 0$,
   $Z - Z' \to 0$,
   $Z - Z'' \to 0$.

Output: E, W, Z.

the reconstructed source data is $\{(\mathbf{x}_i^R, y_i^R \in \{1, 2\})\}_{i=1}^{N_R}$. Our goal here is to learn an effective classifier in the target domain by utilizing both the target data as well as the discriminately reconstructed source data.

To avoid the negative transfer problem, we adopt the principle of *classifier sharing* [49], [53], [54]. Based on it, we develop dual-domain boosting, which discovers and delivers truly useful reconstructed source knowledge to target domain. This learning process can be considered as a multi-task learning problem. Specifically, two boosted classifiers are learned simultaneously in our approach. One classifier is for the target data, while the other is for the reconstructed source data. To leverage the abundant data in the source to regularize the learning process in the target, the principle of classifier sharing assumes that the two boosted classifiers share commonly selected weak learners. On the other hand, the two classifiers have their own ensemble coefficients of the weak learners. This property is used to model the underlying differences between the two domains. In the following, the construction of weak learner candidates is firstly introduced, and the algorithm of dual-domain boosting is then described.

1) Weak learner construction: We employ the RBF kernel function to measure the similarity between data from the target dataset and the reconstructed source dataset. That is,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\gamma^2}\Big), \tag{21}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ can be target and reconstructed source data, and γ is a positive constant. We empirically set γ as the average of the pairwise distances.

Algorithm 2: *Dual-domain Boosting*

Input: reconstructed source data $\{(\mathbf{x}_i^R, y_i^R \in \{1, 2\})\}_{n=1}^{N_R}$, target data $\{(\mathbf{x}_i^T, y_i^T \in \{1, 2\})\}_{n=1}^{N_T}$, iteration V.

Output: source classifier f^R and target classifier f^T, where
$f^\ell(\mathbf{x}) = \sum_{t=1}^{V} \beta_t^\ell h_t(\mathbf{x})$ for $\ell \in \{R, T\}$.

Initialize: $w_1^\ell = w_2^\ell = \cdots = w_{N_\ell}^\ell = \frac{1}{N_\ell}$ for $\ell \in \{R, T\}$.

for $t \leftarrow 1, 2, \ldots, V$ **do**

1. Select the optimal dyadic hypercut h_t by
   $h_t = \arg\min_h \sum_{\ell \in \{R,T\}} \sum_{i=1}^{N_\ell} w_i^\ell [\![h_t(\mathbf{x}_i^\ell) \neq y_i^\ell]\!]$.
2. Compute coefficient $\beta_t^\ell = \max\left(0, \frac{1}{2} \ln \frac{1-\epsilon_\ell}{\epsilon_\ell}\right)$ for $\ell \in \{R, T\}$ where
   $\epsilon_\ell = \sum_{i=1}^{N_\ell} w_i^\ell [\![h_t(\mathbf{x}_i^\ell) \neq y_i^\ell]\!]$.
3. Update data weights $\{w_i^\ell\}_{i=1}^{N_\ell}$ for $\ell \in \{R, T\}$ by
   $w_n^\ell = w_n^\ell \exp\left(-2\beta_t^\ell [\![h_t(\mathbf{x}_i^\ell) \neq y_i^\ell]\!]\right)$.
4. Normalize data weights $\{w_i^\ell\}_{i=1}^{N_\ell}$ for $\ell \in \{R, T\}$.

To learn a boosted classifier with kernel function k, we adopt the method proposed in [55]. The discriminant power of a kernel is first converted into a set of weak learners, called *dyadic hypercuts*. It turns out that the transfer learning can be achieved by boosting over the pool of dyadic hypercuts yielded from the kernel function.

A dyadic hypercut h is composed of two elements here, including a pair of samples, $\mathbf{x}_p$ and $\mathbf{x}_n$, from the opposite classes, i.e., $y_p = 1$ and $y_n = 2$. Note that $\mathbf{x}_p$ and $\mathbf{x}_n$ can be from the target or the reconstructed source domains. The yielded dyadic hypercut h is

$$h(\mathbf{x}) = \begin{cases} 1, & \text{if } k(\mathbf{x}_p, \mathbf{x}) - k(\mathbf{x}_n, \mathbf{x}) - b \geq 0, \\ 2, & \text{otherwise,} \end{cases} \tag{22}$$

where b is a threshold, whose value is determined by error minimization in boosting. The size of the weak learner pool is $|\mathcal{H}| = N^p \times N^n$, where N^p and N^n are the numbers of data of class 1 and 2, respectively.

2) Boosting algorithm: To transfer knowledge from the reconstructed source domain to the target domain, we follow the strategy of classifier sharing in [49], [54] where the relatedness between tasks are modeled by the shared weak learners while the differences between tasks are reflected by their respective ensemble coefficients. The binary classifiers f^R and f^T in the two domains are respectively given by

$$f^\ell(\mathbf{x}) = \sum_{t=1}^{V} \beta_t^\ell h_t(\mathbf{x}), \text{ for } \ell \in \{R, T\}, \tag{23}$$

where $\{h_t\}$ are the shared weak learners, V is the number of the selected weak learners, and $\{\beta_t^\ell\}$ are the respective coefficients.

Algorithm 2 shows a systematic way used to learn f^R and f^T simultaneously. It can be proved that the dyadic hypercut h_t selected in step 1 minimizes the sum of the *exponential loss*

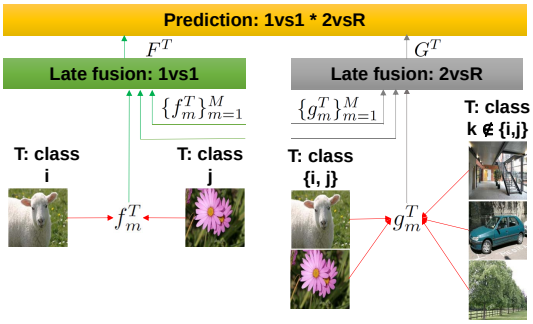


Fig. 2. Late fusion of the learned one-verse-one (1vs1) and one-verse-rest (2vsR) classifiers.

in the two domains, while the ensemble coefficients given in step 2 are determined to minimize the exponential loss of the corresponding domains, respectively. Thus, our approach is developed with the theoretical support of AdaBoost, since the *exponential loss* is monotonically decreased. By sharing weak learners, it turns out that the high risk of overfitting caused by insufficient samples in the target domain can be considerably alleviated, because knowledge extracted from the correlated source domain can regularize the selection of weak learners in the target. Furthermore, the intrinsic differences between the two domains can be modeled through their respective ensemble coefficients whose values are set by considering data in individual domains. Hence, we relieve the unfavorable effect of *negative transfer*, which typically happens if the differences between the source and target domains are not properly addressed during knowledge transfer.

We notice that the performance of using one-verse-one scheme for multi-class classification can be further improved. The observation is that a one-verse-one classifier is learned by using training data of the corresponding two classes in the stage of training. However, this classifier is applied to all testing data, which can be of any classes, in the stage of testing. It is not reasonable, especially in our cases of transfer learning, since all the source data may be diversely transformed and reconstructed in learning these one-verse-one classifiers. To address this issue, we associate each one-verse-one classifier with one additional *two-verse-rest* (2vsR) classifier, which is designed to separate data of the two target categories from the rest target data. In our implementation, the *two-verse-rest* classifier is also a boosted classifier which combines dyadic hypercuts, and it is derived with only target domain data. That is, $\mathbf{x}_p$ in Eq. (22) can be a data sample of target classes 1 and 2, while $\mathbf{x}_n$ in Eq. (22) can be one sample of the rest classes. With the pool of dyadic hypercuts generated in this way, AdaBoost is employed to learn the two-verse-rest classifier. The yielded classifier is denoted by g^T. In sum, a one-verse-one classifier f^T coupled with a two-verse-rest classifier g^T is derived for data in the target domain.

E. Multiple Feature Combination via Late Fusion

Our approach can work on data in multiple feature representations. Given M feature descriptors, we repeat the aforementioned procedure of classifier learning, and have a set of one-verse-one classifiers $\{f_m^T\}_{m=1}^M$ and a set of two-verse-rest classifiers $\{g_m^T\}_{m=1}^M$, i.e., one pair of classifiers for each feature representation.

Late fusion is adopted for feature combination in this work. Specifically, one-verse-one classifiers $\{f_m^T\}_{m=1}^M$ are firstly normalized by applying the sigmoid function to the outputs of these classifiers. Then, the normalized outputs are concatenated, and an SVM classifier with the RBF kernel is derived to work on the concatenated vectors. In this way, features are combined in the classifier level. The same procedure for feature combination is applied to two-verse-rest classifiers $\{g_m^T\}_{m=1}^M$, except the SVM classifier is derived with probability outputs. After late fusion, the two resulting classifiers are denoted by F^T and G^T, respectively. For multi-class classification with the one-verse-one strategy, the output of F^T is weighted by that of G^T in voting. The class with the maximal value in voting is then predicted.

Fig. 2 illustrates how late fusion works for feature combination in this work. For each pair of target classes and each feature m, a one-verse-one classifier f_m^T and a one-verse-rest classifier g_m^T are firstly learned. Late fusion then performs over $\{f_m^T\}_{m=1}^M$ and $\{g_m^T\}_{m=1}^M$ for feature combination, and yields fused classifiers F^T and G^T, respectively. The output of F^T is weighted by that of G^T in final prediction.

IV. Experimental Results

The proposed method is evaluated in this section. We describe in turn the adopted datasets, features, approaches for comparison, the within-database and cross-database transfer learning settings along with their results, and the effects of using late fusion and two-verse-rest classifiers.

A. Datasets

To evaluate the performance of the proposed method, we conduct experiments on three publicly available datasets, Caltech256 [20], SUN09 [21], and MSRC [22]. Although each of these datasets has its own emphasis [56], they are popular benchmarks of object recognition due to their broad coverage of object characteristics and divergent appearances of objects within a single category.

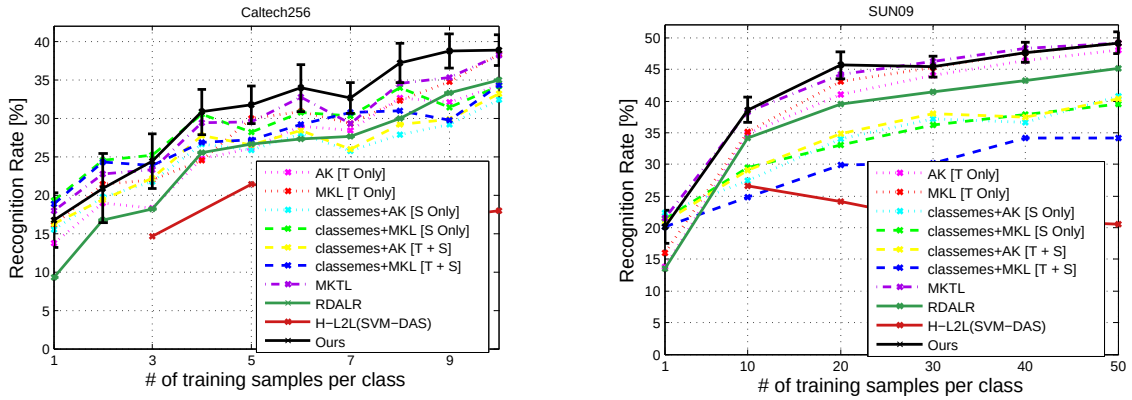
B. Features

Generally speaking, there is no universal feature that can be effectively used to recognize diverse object categories. Hence, we select four representative features to capture various characteristics of images, including:

GIST: We apply the *gist* descriptor [57] to the resized images with a 128×128 pixel prior. For normalization, we firstly compute the mean and the standard deviation of all training data in each feature dimension. For both training and unseen testing data, normalization is then performed by using the computed mean and standard deviation of training data.

BoW-SIFT: We randomly sample interest points detected in an image and describe them by the SIFT descriptor [58]. With a dictionary of $1,000$ visual words, each image is represented as a histogram using this dictionary.

Color histogram: We use a 166-bin color histogram extracted from the HSV color space to represent an image, where



(a) Caltech256 (b) SUN09

Fig. 3. Within-database transfer learning on (a) Caltech256 dataset and (b) SUN09 datasets. Recognition rates of all the evaluated approaches with different numbers of training data are plotted.

the hue, saturation, and intensity channels are divided into 18, 3, and 3 bins, respectively, and there are four additional scales of intensity for describing gray images.

Texton: We consider 99 filters from three filter banks to generate the vocabularies of texture prototypes [59]. Hence, an image can be represented by a histogram that records its probability distribution over all the generated textons.

C. Baselines

Our approach is compared with a few baselines and the state-of-the-art approaches to transfer learning. Based on the information used for model learning, we roughly divide these methods for comparison into the following three categories:

Target Only: Neither auxiliary source data nor prior knowledge are involved in the baselines of this category. Since kernel methods working with multiple features are powerful methodologies for object recognition, we adopt two multi-kernel extensions of SVMs, including the *average kernel* (AK) suggested in [5] and the ensemble kernel learned by the *multiple kernel learning* (MKL) software [4]. The two baselines are respectively denoted as *AK [T Only]* and *MKL [T Only]*. In the two methods, data under each adopted feature are represented as a kernel, and the resulting four kernel matrices then serve as the input.

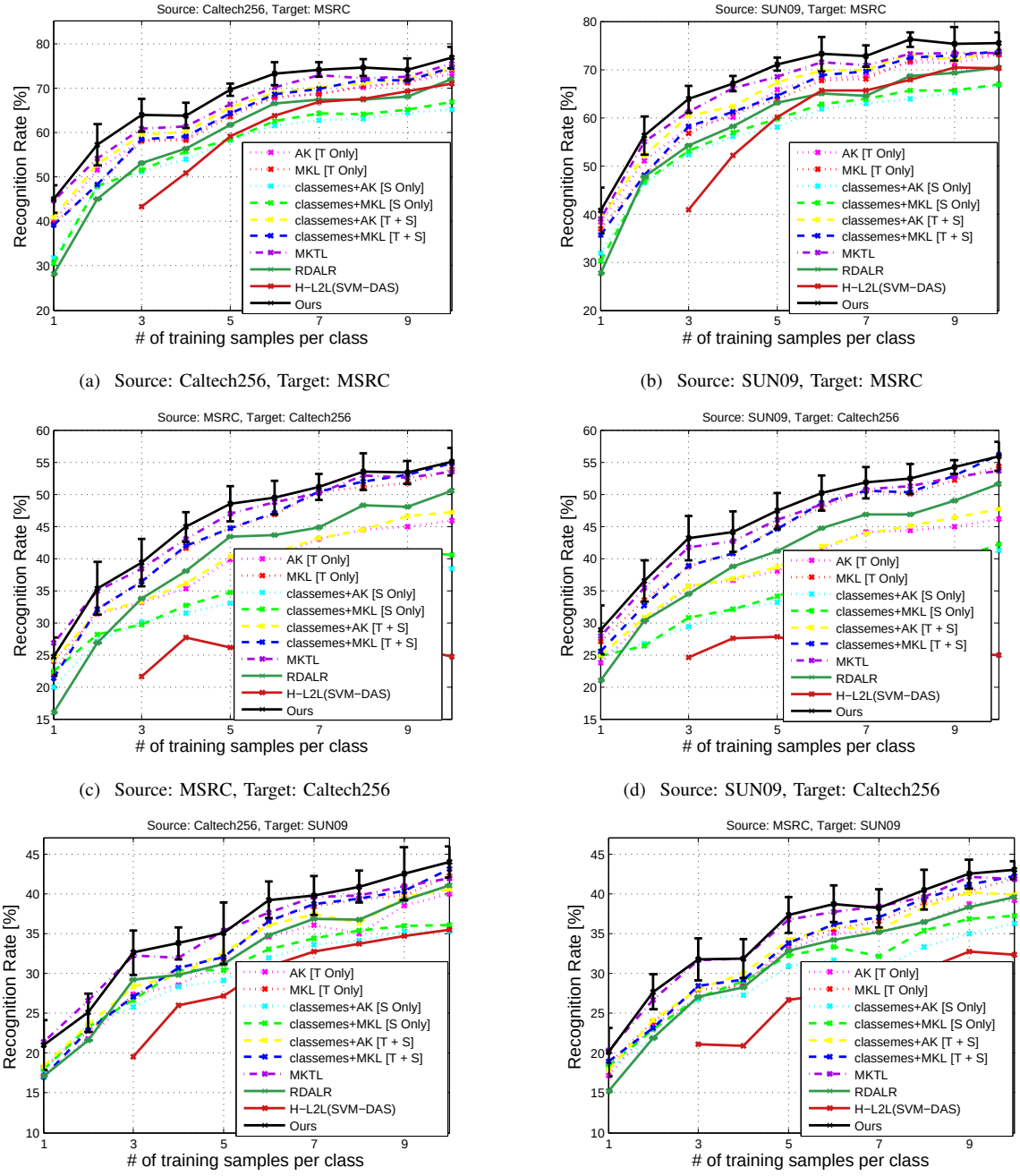
Source Only: We implement the *classemes* [44], a set of powerful features designed for transfer learning, to deliver knowledge from source to target. To this end, an SVM-based classifier with probabilistic outputs is learned for each source object class. The classemes are the probabilistic estimates obtained by applying these classifiers to the target data. While the procedure is performed for each of the adopted features, four new kernels based on classemes are constructed. Baselines *classemes+AK [S Only]* and *classemes+MKL [S Only]* are then respectively established by coupling AK and MKL to the four new kernels.

Target + Source: To fuse information from both the two domains, we jointly consider the kernels yielded by visual features and classemes. Similarly, baselines *classemes+AK [T+S]* and *classemes+MKL [T+S]* are established. Besides, our approach is compared with *multi kernel transfer learning* (MKTL) [46] and *High level-Learning2Learn* (H-L2L) [60], which are two of the best transfer learning algorithms and support heterogeneous transfer from different kinds of priors. For H-L2L, there are two ways of implementation, including *SVM-DAS* and *LP-β*. We implemented the former, termed H-L2L (SVM-DAS), where the output confidences from target and source domains are augmented into a new feature representation. Refer to [60] for the details. Besides, we adopt one-verse-one strategy for multi-class classification and late fusion, and implement *robust domain adaptation with low-rank reconstruction* (RDALR) [40] for comparison. Our approach is established upon RDALR for reconstruction-based transfer learning. Our approach further carries out the reconstruction of source data in a discriminative manner.

D. Within-database transfer learning

We first carry out within-database transfer learning on SUN09 and Caltech256 datasets. In this setting, data in the source and target domains come from the same dataset. Thus, the collection setup and the imaging style of the source and target data are similar, but the source and target categories are completely different. For Caltech256 dataset, 30 target categories and 90 source categories are randomly selected. We pick 1 to 10 samples as well as 50 samples from each target class for training and testing respectively. For SUN09 dataset, 10 target categories and 20 source categories are randomly selected. We pick 1 to 50 samples as well as 50 samples from each target class for training and testing respectively. For both datasets, the number of samples in each source class is set as 50. All experiments are repeated ten times to reduce the effect of sampling.

For parameter tuning, α in Eq. (13) controls the relative importance of the reconstruction errors. In the experiment, we conducted two sets of within-database transfer learning here, and six sets of cross-database transfer learning in the next section. Each experiment set is repeated ten times by using different splits of training and testing data. We set the value of α as $\{10^{-3}, 10^{-2}, ..., 10^{3}\}$, and chose the optimal value of α for each experiment set. The optimal value of α is 10^{-1} for all the six experiment sets of cross-database transfer



(a) Source: Caltech256, Target: MSRC

(b) Source: SUN09, Target: MSRC

(c) Source: MSRC, Target: Caltech256

(d) Source: SUN09, Target: Caltech256

(e) Source: Caltech256, Target: SUN09

(f) Source: MSRC, Target: SUN09

Fig. 4. Cross-database transfer learning on Caltech256, MSRC and SUN09 datasets. (a) $\sim$ (f) Six source-target combinations on the three datasets. Recognition rates are plotted as a function of the number of training samples per class.

learning. The optimal values of α vary in the two sets of within-database transfer learning. The performance of dual-domain boosting in Algorithm 2 typically converges when the number of weak learners, V, is more than 10. We set V as 25 in all the experiments.

In Fig. 3, we plot the average recognition rates of our approach and other methods. For the sake of clearness, only the standard deviation of our approach is shown. The consistent trend of curves in both two datasets indicates that our approach outperforms the other baselines mostly. Although baselines with only source priors, e.g., *classemes+AK [S only]*, don't perform well, incorporating the source knowledge indeed helps in constructing more accurate classifiers, e.g., *classemes+AK [T+S]*. Besides, baselines with multiple kernel learning work better than baselines with average kernel in most of the cases. This is because SimpleMKL [4] can effectively select the better kernel combination to emphasize discriminant features for the given data.

On the other hand, compared with RDALR [40], our method achieves better recognition accuracy. The results confirm the advantage of using both discriminant reconstruction and two-verse-rest classifiers. With regard to MKTL, it works better

TABLE I
THE PERFORMANCE $[acc \pm std\%]$ OF DIFFERENT APPROACHES ON EIGHT SOURCE-TARGET COMBINATIONS WHEN THE NUMBER OF TARGET TRAINING DATA PER CLASS IS SET AS TEN.

Type	Within-database		Cross-database					
Tartget dataset	SUN09	Caltech256	MSRC		Caltech256		SUN09	
Source dataset	SUN09	Caltech256	Caltech256	SUN09	MSRC	SUN09	Caltech256	MSRC
AK [T Only]	34.46±0.94%	34.27±1.04%	73.16±2.61%	73.04±2.51%	45.84±2.76%	46.14±1.88%	40.02±2.44%	39.24±2.06%
MKL [T Only]	35.08±2.37%	38.30±2.22%	74.28±3.19%	73.02±2.37%	53.84±2.62%	54.36±2.21%	42.26±2.45%	41.96±2.57%
classemes+AK [S Only]	28.12±3.03%	32.37±0.33%	65.06±1.81%	66.74±1.69%	38.34±2.37%	41.24±1.94%	35.26±2.41%	36.32±1.71%
classemes+MKL [S Only]	29.58±3.12%	34.20±0.19%	66.84±1.41%	66.84±1.77%	40.58±2.71%	42.24±3.05%	36.08±1.82%	37.28±1.94%
classemes+AK [T + S]	28.54±2.46%	33.20±0.38%	74.38±2.37%	74.00±2.62%	47.26±2.80%	47.64±1.51%	40.62±2.73%	39.92±1.87%
classemes+MKL [T + S]	26.04±4.20%	34.27±0.19%	74.62±2.90%	73.72±2.04%	54.84±1.48%	**56.16±1.71%**	43.12±2.58%	42.24±2.22%
H-L2L(SVM-DAS)	26.56±1.95%	17.93±7.07%	70.92±4.27%	70.30±2.79%	24.68±3.43%	24.96±2.78%	35.52±2.43%	32.32±3.55%
MKTL [46]	38.18±1.76%	38.20±1.51%	75.46±3.00%	73.42±1.96%	53.56±2.23%	53.64±1.20%	41.94±2.15%	41.82±1.71%
RDALR [40]	34.17±2.08%	35.00±2.00%	73.77±1.74%	70.68±2.82%	50.59±3.50%	52.40±2.25%	41.51±2.56%	40.14±1.55%
Ours	**38.64±1.98%**	**38.87±2.00%**	**76.83±2.39%**	**75.47±2.22%**	**55.04±2.17%**	55.89±2.22%	**43.98±1.96%**	**43.01±1.15%**

than all other baselines. In within-database transfer learning, our approach obtains slightly better performance than MKTL. Nevertheless, we will show that the performance gain of our method over MKTL will become notable in cross-database transfer learning, where the variations between source and target data are larger, and correlating the two domains in advance is hence more crucial.

E. Cross-database transfer learning

We then evaluate the proposed approach for cross-database transfer learning. The data in the source domain and the target domain are from different databases. With three datasets Caltech256, MSRC, and SUN09, totally we have six source-target combinations. For each source-target combination, 10 target categories and 20 source categories are randomly selected. The number of samples in each source class is set as 50. The number of training data per target class varies from 1 to 10, while the number of testing data is fixed as 50 per class. All experiments are repeated ten times to reduce the effect of sampling.

In Fig. 4, the recognition rates of our approach as well as all approaches for comparison are shown. In the six settings of cross-database of transfer learning, our approach achieves the best results in most cases, similar to what we have observed in within-database transfer learning. Compared with RDALR [40], which transfers knowledge by low-rank reconstruction only, our method with the proposed discriminative term and two-verse-rest classifiers achieves remarkable performance improvement. In cross-database transfer learning, the advantage of our approach becomes more obvious owning to the larger inter-database variations. When Caltech256 and MSRC datasets yield the source-target combination, our method has considerable performance gains over MKTL, which is one of the state-of-the-art methods for transfer learning. The main reason is that objects in the two datasets are too dissimilar to directly transfer knowledge across them. Correlating the two domains via adapting the inter-database variations is helpful for the successive knowledge transfer. On the other hand, our method and MKTL achieve very similar performance, when transferring knowledge from MSRC to SUN09, because variations between the two datasets are smaller.

We report the mean accuracy and the standard deviation of our approach and the competing approaches in TABLE I. Both the mean and the standard deviation of each method in eight sets of experiments, including two sets of within-database transfer learning and six sets of cross-database transfer learning, are given in the case where the number of samples per target class is set as 10. Note that using few randomly selected training data leads to the large performance variations of each approach. Our approach in most settings give the best mean accuracy.

Our approach is compared with nine competing approaches in this work. Among them, MKTL on average gives the best performance. For significance test, we report the *relative improvement* of our approach with respect to MKTL on the eight experiment sets in TABLE II. Each entry in this table is computed via $\frac{(A-B)}{B} \times 100\%$ where A and B are the mean accuracy rates of our approach and MTKL, respectively. Compared with MKTL, our approach gets the comparable performance for within-database transfer learning, and is superior to MKTL for cross-database transfer learning. The results show that our approach can work relatively well when the correlation between the source and target domains is weak.

We have conducted the experiments with two main settings, i.e., within-database transfer learning and cross-database transfer learning, for comparing our method with other methods. The correlation between source and target domains is higher in the former setting, while it is lower in the latter setting. Comparing the results in Fig. 3, Fig. 4, and TABLE I, it can be found that our approach shows its superiority to the completing methods in the latter setting. In the setting where higher correlation between source and target domains exists, our approach and the state-of-the-art method, MKTL [46], give similar performance, but our approach needs the extra computational cost for source data reconstruction.

F. Running time

The efficiency of the compared approaches is evaluated here. We consider the setting where within-database transfer learning is conducted on SUN09 dataset and the number of target training data per class is set as 10. The training time of our approach and the competing approaches on a modern PC with an Intel Core $i7$ 3.4 GHz processor is given in TABLE III.

TABLE II
THE RELATIVE IMPROVEMENT (%) OF OUR APPROACH W.R.T. MKTL.

Type	Within-database		Cross-database					
Tartget dataset	SUN09	Caltech256	MSRC		Caltech256		SUN09	
Source dataset	SUN09	Caltech256	Caltech256	SUN09	MSRC	SUN09	Caltech256	MSRC
Ours vs. MKTL	1.20%	1.75%	1.81%	2.79%	2.76%	4.19%	4.86%	2.84%

TABLE III
TRAINING TIME OF VARIOUS APPROACHES FOR WITHIN-DATABASE TRANSFER LEARNING ON SUN09 DATASET WITH TEN TRAINING SAMPLES PER TARGET CATEGORY.

Method	Time (seconds)
AK [T Only]	1.46
MKL [T Only]	1.41
classemes+AK [S Only]	41.66
classemes+MKL [S Only]	397.88
classemes+AK [T + S]	43.24
classemes+MKL [T + S]	730.36
H-L2L (SVM-DAS)	682.41
MKTL [46]	41.3
RDALR [40]	627.66
Ours	1045.01

The proposed approach is implemented in `Matlab`, except that support vector machines (SVMs) which are used in late fusion and is implemented in `C`. Approaches, including MKL, classemes+MKL [S Only], classemes+MKL [T + S], RDALR, and H-L2L (SVM-DAS), are mainly implemented in `Matlab`, while the rest are mainly implemented in `C`. Some MKL-based approaches, e.g., classemes+MKL [T + S], and reconstruction-based approaches, e.g., RDALR and ours, are less efficient in training.

G. Effects of using late fusion and two-verse-rest classifier

In this work, we adopt multiple feature representations for better data description, and combine them via late fusion. We hence evaluate the effectiveness of the fusion process in our approach. Our approach can work with a single feature representation and multiple ones. For fair evaluation, our approach to training the one-verse-one classifiers is repeatedly applied to the four adopted representations five times, one for each representation and one for the four representations jointly. On the other hand, we also aim to measure the effect of using the two-verse-rest classifiers, especially when abundant source data are transformed and included in the training process. Thus, we compare the recognition accuracy with and without using the two-verse-rest classifiers in the cases where the four representations are jointly used.

Fig. 5 shows the results in cross-database transfer learning. We can observe that late fusion is indeed helpful in improving the performance. It is understood that a single feature representation does not suffice for recognizing objects over diverse object categories and on different datasets. Late fusion in our approach effectively leverages the four complementary feature representations to consistently boost the recognition rates. It is worth mentioning that the advantage of using the two-verse-rest classifiers is very significant in our cases. We investigate into the effect, and find that the source data are diversely transformed and reconstructed in learning each of the one-verse-one classifiers. The numbers of source data reconstructed by the two target classes are probably different. The unbalanced training data makes the learned one-verse-one classifier tend to have bias towards some target class for test data of the irrelevant classes. The developed two-verse-rest classifiers alleviate this problem, and hence improve the accuracy.

H. Visualization

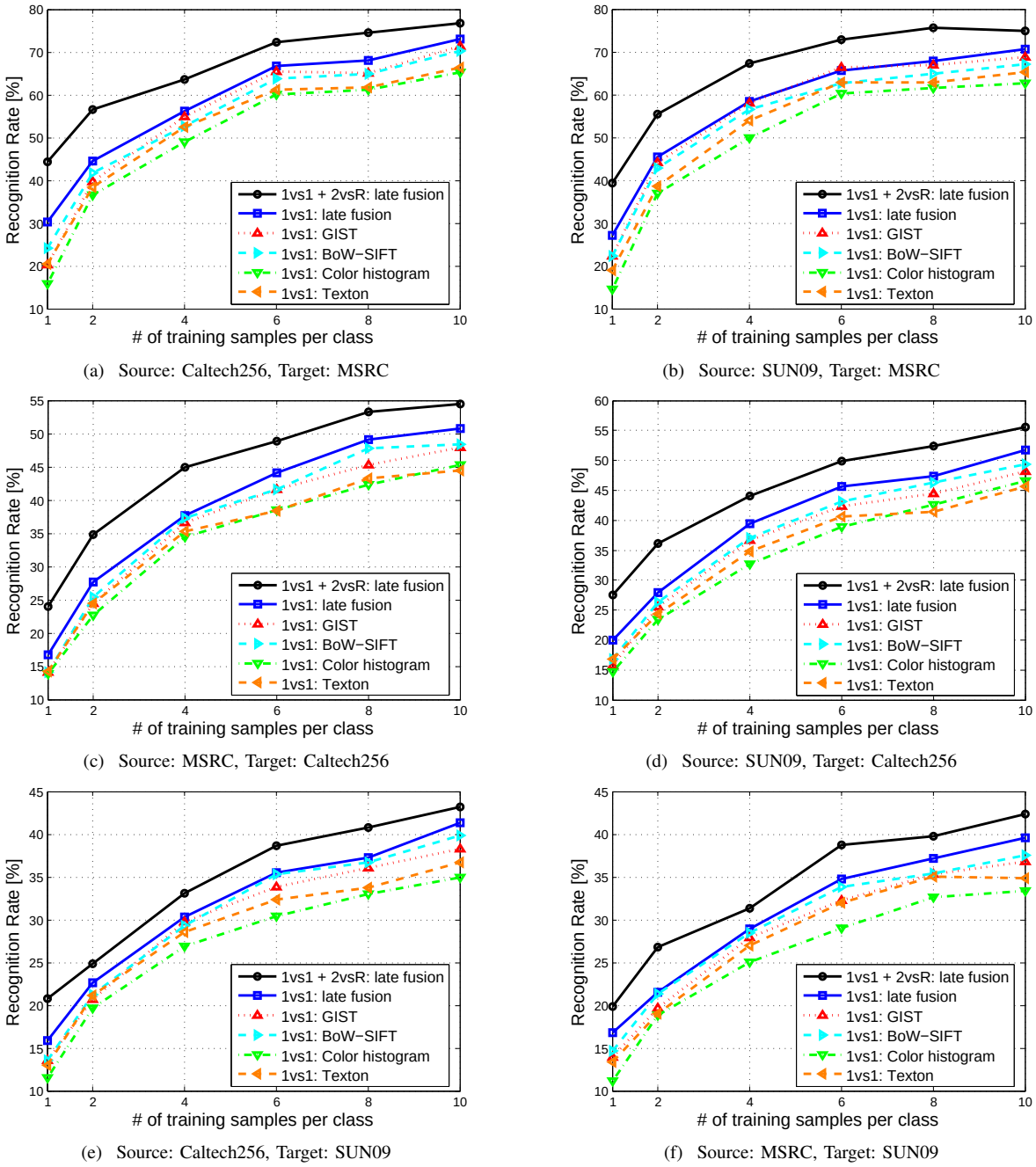
To gain insight into the quantitative results, we *visualize* how the proposed discriminative reconstruction behaviors for source class partition. Fig. 6 shows eight examples of how the source classes are divided into the two target classes, each example in one row of this figure. There are four feature representations are considered, and two examples are picked for each representation. All the eight examples are selected, when Caltech256 and MSRC datasets serve as the source and the target datasets, respectively. It is due to that the inter-database variations are large between the two datasets. Consider the first example, i.e., the first row of Fig. 6. The one-verse-one classifier is derived for target classes `buildings` and `scenes-urban`. Sample images of the two classes are shown in columns (a) and (h) respectively. Through discriminative reconstruction in Algorithm 1, the top three source classes, i.e., those with the lowest reconstruction errors, that are assigned to target class `buildings` are shown in columns (b) $\sim$ (d), respectively. Similarly, the top three source classes assigned to `scenes-urban` are given in columns (g) $\sim$ (e), respectively.

As can be seen in the first two examples, each target class is augmented with the source classes with high similarity in terms of spatial texture layout, which is characterized by GIST. In the fifth and sixth examples where color histogram is adopted as the data representation, each target class and its augmented source classes have similar color distributions. The similarity between the target and source classes in terms of local texture can be found in the rest four examples.

The results in Fig. 6 demonstrate that the proposed discriminative reconstruction can effectively correlate the two domains, and associate each target class with appropriate source classes faithfully based on the adopted feature representation. We consider that it is the reason why the successive dual-domain boosting can achieve successful knowledge transfer across different domains. It is also worth pointing out that our approach works in an *interpretable* way, since how the sources classes are partitioned is known, as those shown in Fig. 6.

V. CONCLUSIONS

We have presented an effective approach that can leverage useful knowledge in the source domain to facilitate classifier



(a) Source: Caltech256, Target: MSRC

(b) Source: SUN09, Target: MSRC

(c) Source: MSRC, Target: Caltech256

(d) Source: SUN09, Target: Caltech256

(e) Source: Caltech256, Target: SUN09

(f) Source: MSRC, Target: SUN09

Fig. 5. Effects of using late fusion and two-verse-rest classifiers in the six settings of cross-database transfer learning.

learning in the target domain, especially when few training examples in target are available. Our approach makes no assumption about the correlation between the source and target domains, since it can correlate the two domains via reconstructing source data by target data in a discriminative manner. During the process of reconstruction, the source data are firstly transformed to adapt inter-domain variations, and only the reconstructed part is borrowed to enrich the corresponding training set in target. The developed dual-domain boosting then casts knowledge transfer as a multi-task learning problem. It derives the boosted classifier in target, and regularizes it via joint weak learner selection for both the target and the reconstructed source data. In addition, we have presented the two-verse-rest classifiers, which alleviate the problem in the one-verse-one voting strategy for multi-class classification, and improve the performance. The proposed approach is comprehensively evaluated on three benchmark datasets of object recognition with both the settings of within- and cross-database transfer. The effects of using the proposed components, including discriminative reconstruction, feature fusion, and two-verse-rest classifiers, have been evaluated. Both the quantitative and visualization results demonstrate that our approach can significantly improve the performance of the learned classifiers in target by making the most of the abundant data in source.

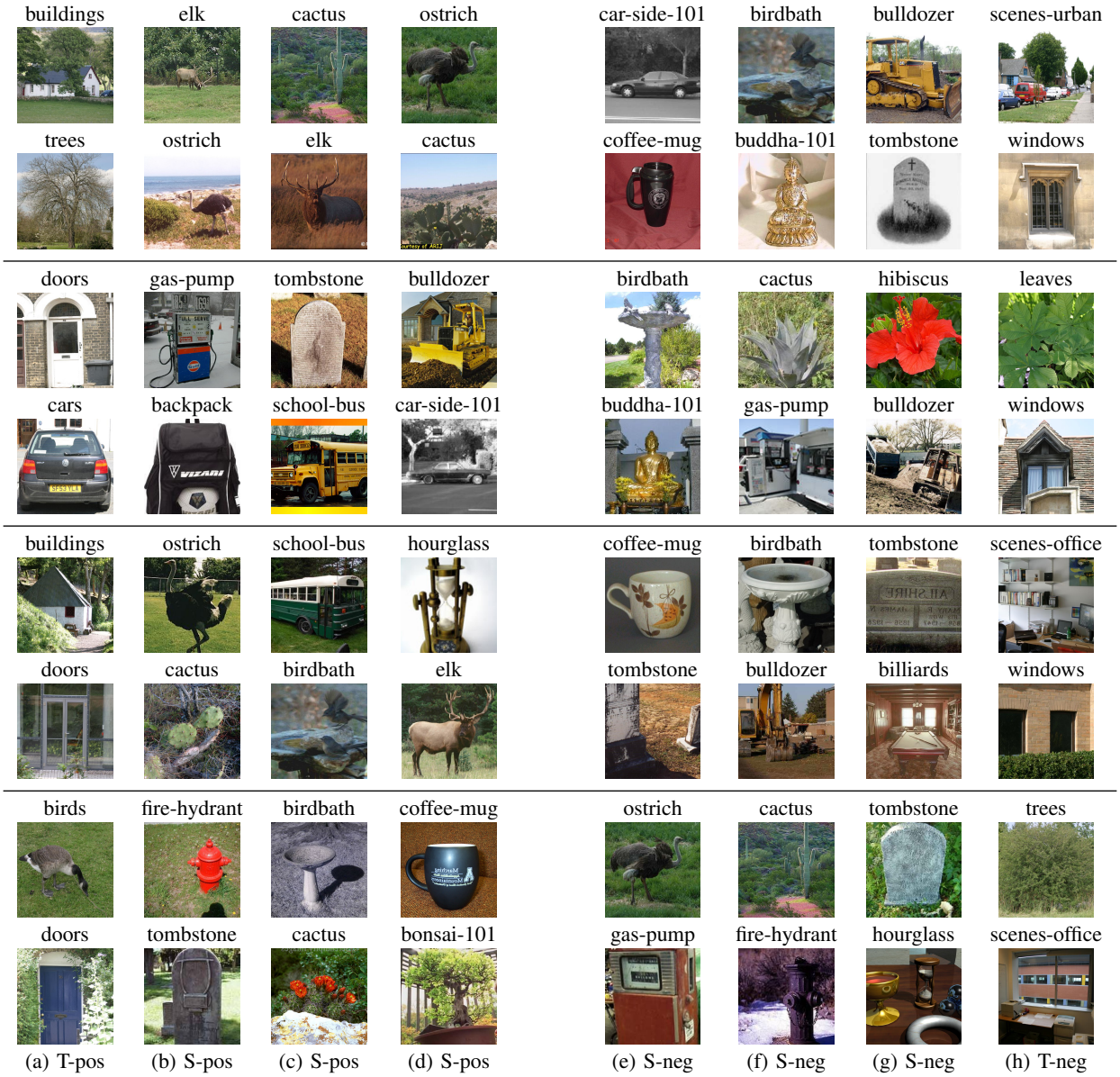


Fig. 6. Eight visualization examples of the proposed discriminative reconstruction, each in one row. Two examples for each adopted feature representations, i.e., GIST (rows 1 & 2), BoW-SIFT (rows 3 & 4), Color histogram (rows 5 & 6), and Texton (rows 7 & 8). (a) & (h) Two target classes with their sample images in learning the one-verse-one classifier. (b) $\sim$ (d) Top three source classes assigned to target class in (a). (g) $\sim$ (e) Top three source classes assigned to target class in (h). See the text for the details.

REFERENCES

[1] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. Int'l Conf. Machine Learning*, Jul. 2004, pp. 6–13.

[2] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Machine Learning Research*, vol. 5, pp. 27–72, Dec. 2004.

[3] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. Int'l Conf. Computer Vision*, Oct. 2007, pp. 1–8.

[4] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[5] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. Int'l Conf. Computer Vision*, Oct. 2009, pp. 221–228.

[6] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1147–1160, June 2011.

[7] K.-H. Liu, Y.-Y. Lin, and C.-S. Chen, "Linear spectral mixture analysis via multiple kernel learning for hypercpectral image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2254–2269, Apr. 2015.

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

[9] Z. Guo and Z. J. Wang, "Cross-domain object recognition via input-output kernel analysis," *IEEE Trans. on Image Processing*, vol. 22, no. 8, pp. 3108–3119, Aug. 2013.

[10] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. Int'l Conf. Machine Learning*, Jun. 2007, pp. 193–200.

[11] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[12] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *Proc. ACM Conf. Multimedia*, Sep. 2007, pp. 188–197.

[13] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2010, pp. 1855–1862.

[14] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1717 –1724.

[15] J. Kim and K. Grauman, "Shape sharing for object segmentation," in *Proc. Euro. Conf. Computer Vision*, Oct. 2012, pp. 444 –458.

[16] Q. Wang, F. Chen, J. Yang, W. Xu, and M.-H. Yang, "Transferring visual prior for online object tracking," *IEEE Trans. on Image Processing*, vol. 21, no. 7, pp. 3296–3305, Jul. 2012.

[17] G. Chen, C. Xiong, and J. J. Corso, "Dictionary transfer for image denoising via domain adaptation," in *Proc. Int'l Conf. Image Processing*, Sep. 2012, pp. 1189–1192.

[18] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Euro. Conf. Computer Vision*, Sep. 2010, pp. 213–226.

[19] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: domain adaptation using asymmetric kernel transforms," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2011, pp. 1785–1792.

[20] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep., 2007.

[21] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2010, pp. 3485–3492.

[22] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. Int'l Conf. Computer Vision*, vol. 2, Oct. 2005, pp. 1800–1807.

[23] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 26, pp. 1019–1034, May 2015.

[24] J. Huang, A. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," in *Advances in Neural Information Processing Systems*, Dec. 2007, pp. 601–608.

[25] L. Duan, I. Tsang, D. Xu, and S. Maybank, "Domain transfer SVM for video concept detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1375–1381.

[26] T. Tommasi, F. Orabona, and B. Caputo, "Safety in numbers: Learning categories from few examples with multi model knowledge transfer," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2010, pp. 3081–3088.

[27] W. Jiang, E. Zavesky, S.-F. Chang, and A. Loui, "Cross-domain learning methods for high-level visual concept classification," in *Proc. Int'l Conf. Image Processing*, Oct. 2008, pp. 161–164.

[28] Y. Luo, T. Liu, D. Tao, and C. Xu, "Decomposition-based transfer distance metric learning for image classification," *IEEE Trans. on Image Processing*, vol. 23, pp. 3789–3801, Sep. 2014.

[29] Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo, "Domain adaptive semantic diffusion for large scale context-based video annotation," in *Proc. Int'l Conf. Computer Vision*, Sep. 2009, pp. 1420–1427.

[30] G.-J. Qi, C. Aggarwal, Y. Rui, Q. Tian, S. Chang, and T. Huang, "Towards cross-category knowledge propagation for learning visual concepts," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2011, pp. 897–904.

[31] F.-J. Chang, Y.-Y. Lin, and M.-F. Weng, "Cross-database transfer learning via learnable and discriminant error-correcting output codes," in *Proc. Asian Conf. on Computer Vision*, Nov. 2012, pp. 16–30.

[32] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1778–1785.

[33] S. J. Hwang, F. Sha, and K. Grauman, "Sharing features between objects and their attributes," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2011, pp. 1761–1768.

[34] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by betweenclass attribute transfer," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2009, pp. 951–958.

[35] D. Mahajan, S. Sellamanickam, and V. Nair, "A joint learning framework for attribute models and object descriptions," in *Proc. Int'l Conf. Computer Vision*, Nov. 2011, pp. 1227–1234.

[36] D. Parikh and K. Grauman, "Relative attributes," in *Proc. Int'l Conf. Computer Vision*, Nov. 2011, pp. 503–510.

[37] M. Rastegari, A. Farhadi, and D. Forsyth, "Attribute discovery via predictable discrimnative binary codes," in *Proc. Euro. Conf. Computer Vision*, Oct. 2012, pp. 876–889.

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

[39] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. Nat'l Conf. Artificial Intelligence*, Jul. 2008, pp. 677–682.

[40] I.-H. Jhuo, D. Liu, D. T. Lee, and S.-F. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2168–2175.

[41] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. Int'l Conf. Computer Vision*, Dec. 2013, pp. 2960–2967.

[42] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1410–1417.

[43] T. Yao, Y. Pan, C.-W. Ngo, H. Li, and T. Mei, "Semi-supervised domain adaptation with subspace learning for visual recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2015, pp. 2142–2150.

[44] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. Euro. Conf. Computer Vision*, Sep. 2010, pp. 776–789.

[45] H. Daume III, "Frustratingly easy domain adaptation," in *Proc. Annual Meeting of the Association of Computational Linguistics*, Jun. 2007, pp. 256–263.

[46] L. Jie, T. Tommasi, and B. Caputo, "Multiclass transfer learning from unconstrained priors," in *Proc. Int'l Conf. Computer Vision*, Nov. 2011, pp. 1863–1870.

[47] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *Int. J. Computer Vision*, vol. 109, pp. 42–59, Aug. 2014.

[48] F. Zhu, L. Shao, and J. Tang, "Boosted cross-domain categorization," in *Proc. British Conf. Machine Vision*, 2014.

[49] Y.-Y. Lin, J.-F. Tsai, and T.-L. Liu, "Efficient discriminative local learning for object recognition," in *Proc. Int'l Conf. Computer Vision*, Oct. 2009, pp. 598–605.

[50] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," UIUC, Tech. Rep., 2009.

[51] J.-F. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optimization*, vol. 20, pp. 1956–1982, Mar. 2010.

[52] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2,1-norm minimization," in *Ann. Conf. Uncertainty in Artificial Intelligence*, Jun. 2009, pp. 339–348.

[53] A. Torralba, K. Murphy, and W. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 854–869, Mar. 2007.

[54] R. Yan, J. Tesic, and J. Smith, "Model-shared subspace boosting for multi-label classification," in *Proc. ACM Conf. Knowledge Discovery and Data Mining*, Aug. 2007, pp. 834–843.

[55] B. Moghaddam and G. Shakhnarovich, "Boosted dyadic kernel discriminants," in *Advances in Neural Information Processing Systems*, Nov. 2003, pp. 761–768.

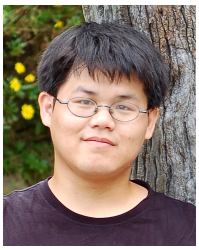
[56] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2011, pp. 1521–1528.

[57] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Computer Vision*, vol. 42, pp. 145–175, May 2001.

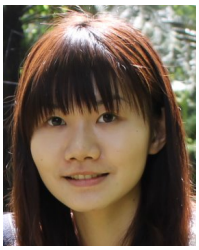
[58] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, pp. 91–110, Nov. 2004.

[59] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Computer Vision*, vol. 62, pp. 61–81, Apr. 2005.

[60] N. Patricia and B. Caputo, "Learning to learn, from transfer learning to domain adaptation: A unifying perspective," in *Proc. Conf. Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1442–1449.



Pai-Heng Hsiao received the B.S. degree in Electrical Engineering in National Tsing Hua University, Taiwan and the M.S. degree in Electronics Engineering from National Taiwan University, in 2007 and 2009, respectively. His current research interests include computer vision, deep learning, machine learning, and intelligent system.



Feng-Ju Chang received her B.S. degree in Electrical Engineering department from National Cheng Kung University and her M.S. degree in Graduate Institute of Communication Engineering from National Taiwan University, in 2009 and 2011 respectively. She was a research assistant in Academia Sinica from 2011 to 2013. She is currently a Ph.D. student in Electrical Engineering department from University of Southern California. Her current research interests include computer vision and machine learning.

Yen-Yu Lin received the B.B.A. degree in information management, and the M.S. and Ph.D. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2001, 2003, and 2010, respectively. He is currently an Associate Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His current research interests include computer vision, pattern recognition, and machine learning. He is a member of the IEEE.