Exploiting Privileged Information from Web Data for Image Categorization

Wen Li*, Li Niu*, and Dong Xu

School of Computer Engineering, Nanyang Technological University, Singapore

Abstract. Relevant and irrelevant web images collected by tag-based image retrieval have been employed as loosely labeled training data for learning SVM classifiers for image categorization by only using the visual features. In this work, we propose a new image categorization method by incorporating the textual features extracted from the surrounding textual descriptions (tags, captions, categories, etc.) as privileged information and simultaneously coping with noise in the loose labels of training web images. When the training and test samples come from different datasets, our proposed method can be further extended to reduce the data distribution mismatch by adding a regularizer based on the Maximum Mean Discrepancy (MMD) criterion. Our comprehensive experiments on three benchmark datasets demonstrate the effectiveness of our proposed methods for image categorization and image retrieval by exploiting privileged information from web data.

Keywords: learning using privileged information, multi-instance learning, domain adaptation.

1 Introduction

Image categorization is a challenging problem in computer vision. A number of labeled training images are often required for learning a robust classifier for image categorization. However, collecting labeled training images based on human annotation is often time-consuming and expensive. Meanwhile, increasingly rich and massive social media data are being posted to the photo sharing websites like Flickr everyday, in which the web images are generally accompanied by valuable contextual information (*e.g.*, tags, captions, and surrounding text). Recently, relevant and irrelevant web images (*e.g.*, Flickr images) collected by tag-based image retrieval have been used as loosely labeled training data for learning SVM classifiers for various computer vision tasks (*e.g.*, image categorization and image retrieval)[43,33,31].

In this work, we extract the visual and textual features from the training web images and the associated textual descriptions (tags, captions, etc.), respectively. While we do not have the textual features in test images, the additional textual features extracted from the training images can still be used as privileged information, as shown in the work [42] from Vapnik and Vashist. Their work is

* Indicates equal contributions.

D. Fleet et al. (Eds.): ECCV 2014, Part V, LNCS 8693, pp. 437–452, 2014.

motivated by human learning, where a teacher provides the students with hidden information through explanations, comments, comparisons etc [42]. Similarly, we observe the surrounding textual descriptions more or less describe the content of training images. So the textual features can additionally provide hidden information for learning robust classifiers by bridging the semantic gap between the low-level visual features and the high-level semantic concepts.

For image categorization using massive web data, another challenging research issue is to cope with noisy labels of relevant training images. To solve this problem, the recent works [43,33,31] partitioned the training images into small subsets. By treating each subset as a "bag" and the images in each bag as "instances", the multi-instance learning (MIL) methods like Sparse MIL (sMIL) [5], mi-SVM [1] and MIL-CPB [33] were used for image categorization and image retrieval.

Based on the above observations, in Section 3, we first propose a new method called *sMIL using privileged information* (sMIL-PI) for image categorization by learning from loosely labeled web data, which not only takes advantage of the additional textual features but also effectively copes with noisy labels of relevant training images. When the training and testing samples are from different datasets, we also observe the data distributions between the training and testing samples may be very different. Our proposed sMIL-PI method can be further extended to reduce the data distribution mismatch. We name the extended method as sMIL-PI-DA, in which we additionally add a regularizer based on the Maximum Mean Discrepancy (MMD) criterion.

In Section 4, we conduct comprehensive experiments for two tasks, image categorization and image retrieval. Our results demonstrate our newly proposed method sMIL-PI outperforms its corresponding existing MIL method (*i.e.*, sMIL), and sMIL-PI is also better than the learning methods using privileged information as well as other related baselines. Moreover, our newly proposed domain adaptation method sMIL-PI-DA achieves the best results when the training and testing samples are from different datasets.

2 Related Work

Researchers have proposed effective methods to employ massive web data for various computer vision applications [37,40,17,27]. Torralba *et al.* [40] used a nearest neighbor (NN) based approach for object and scene recognition by leveraging a large dataset with 80 million tiny images. Fergus *et al.* [17] proposed a topic model based approach for object categorization by exploiting the images retrieved from Google image search, while Hwang and Grauman [27] employed kernel canonical correlation analysis (KCCA) for image retrieval using different features. Recently, Chen *et al.* [6] proposed the NEIL system for automatically labeling instances and extracting the visual relationships.

Our work is more related to [43,11,31,32,33], which explicitly coped with noise in the loose labels of relevant training web images. Those works first partitioned the training images into small subsets. By treating each subset as a "bag" and the images in each bag as "instances", they formulated this task as a multi-instance learning problem. The bag-based MIL method Sparse MIL as well as its variant were used in [43], while an instance-based approach called MIL-CPB was developed in [33]. The works in [43,33] did not consider the additional features in training data, and thus they can only employ the visual features for learning MIL classifiers for image categorization[1]. In contrast, we propose a new image categorization method by incorporating the additional textual features of training images as privileged information.

Our approach is motivated by the work on learning using privileged information (LUPI) [42], in which training data contains additional features (*i.e.*, privileged information) which are not available at the testing stage. Privileged information was also used for distance metric learning [20], multiple task learning [35] and learning to rank [38]. However, all those works only considered the supervised learning scenario using training data with accurate supervision. In contrast, we formulate a new MIL-PI method in order to cope with noise in the loose labels of relevant training web images.

Our work is also related to attributes based approaches [19,15], in which the attribute classifiers are learnt to extract the mid-level features. However, the mid-level features can be extracted from both training and testing images. Similarly, the classeme based approaches [41,30] proposed to use the training images from additionally annotated concepts to obtain the mid-level features. Those methods can be readily applied to our application by using the mid-level features as the main features to replace our current visual features (*i.e.*, the DeCAF features [10] in our experiments). However, the additional textual features, which are not available in the testing images, can still be used as privileged information in our sMIL-PI method. Moreover, those works did not explicitly reduce the distribution mismatch between the training and testing images as in our sMIL-PI-DA method.

Finally, our work is also related to the domain adaptation methods [2,3,18,26,22,21,29,13,4,14,12,34]. Huang *et al.* [26] proposed a two-step approach by re-weighting the source domain samples. For domain adaptation, Kulis *et al.* [29] proposed a metric learning method by learning an asymmetric nonlinear transformation, while Gopalan *et al.* [22] and Gong *et al.* [21] interpolated intermediate domains. SVM based approaches [13,4,14,12] were also developed to reduce the distribution mismatch. Some recent approaches aim to learn a domain invariant subspace [2] or align two subspaces from both domains [18]. Bergamo and Torresani [3] proposed a domain adaptation method which can cope with the loosely labeled training data. However, their method requires the labeled training samples from the target domain which are not required in our domain adaptation method sMIL-PI-DA. Moreover, our sMIL-PI-DA method achieves the best results when the training and testing samples are from different datasets.

[1] The work in [33] used both visual and textual features in the training process. However, it also requires the textual features in the testing process.

3 Multi-Instance Learning Using Privileged Information

Our goal is to learn robust classifiers for image categorization by using automatically collected web images. Given any category name, relevant and irrelevant web images can be collected as training data by using tag-based image retrieval. However, those collected relevant and irrelevant web images may be associated with noisy and inaccurate labels. Moreover, we also observe that web images are usually associated with rich textual descriptions (*e.g.*, tags, captions, and surrounding texts), which provide semantic descriptions to the content of the image to some extent.

To this end, we propose a new learning paradigm called multi-instance learning using privileged information (MIL-PI) for image categorization, in which we not only take advantage of the additional textual descriptions (*i.e.*, privileged information) in training data but also effectively cope with noise in the loose labels of relevant training images. Based on the Sparse MIL (sMIL) method [5], we develop a new method called sMIL-PI in Section 3.2.

When the training and testing samples are from different datasets, the distributions of training and testing samples may be very different. To reduce the data distribution mismatch, we further extend our sMIL-PI method as sMIL-PI-DA for domain adaptation by adding a regularizer based on the Maximum Mean Discrepancy (MMD) criterion into the dual formulation of our sMIL-PI in Section 3.3.

In the remainder of this paper, we use a lowercase/uppercase letter in boldface to denote a vector/matrix (*e.g.*, **a** denotes a vector and **A** denotes a matrix). The superscript $'$ denotes the transpose of a vector or a matrix. We denote $\mathbf{0}_n, \mathbf{1}_n \in \mathbb{R}^n$ as the n-dim column vectors of all zeros and all ones, respectively. For simplicity, we also use $\mathbf{0}$ and $\mathbf{1}$ instead of $\mathbf{0}_n$ and $\mathbf{1}_n$ when the dimension is obvious. Moreover, we use $\mathbf{A} \circ \mathbf{B}$ to denote the element-wise product between two matrices $\mathbf{A}$ and $\mathbf{B}$. The inequality $\mathbf{a} \leq \mathbf{b}$ means that $a_i \leq b_i$ for $i = 1, \ldots, n$.

3.1 Problem Statement

To cope with label noise in the training data, we partition the relevant and irrelevant web images into bags as in the recent works [43,33]. The training bags constructed from relevant images are labeled as positive and those from irrelevant images are labeled as negative.

Formally, let us represent the training data as $\{(\mathcal{B}_l, Y_l)\ |_{l=1}^{L}\}$, where $\mathcal{B}_l$ is a training bag, $Y_l \in \{+1, -1\}$ is the corresponding bag label, and L is the total number of training bags. Each training bag $\mathcal{B}_l$ consists of a number of training instances, *i.e.*, $\mathcal{B}_l = \{(\mathbf{x}_i, \tilde{\mathbf{x}}_i, y_i)|_{i \in \mathcal{I}_l}\}$, where $\mathcal{I}_l$ is the set of indices for the instances inside $\mathcal{B}_l$, $\mathbf{x}_i$ is the visual feature of the i-th sample, $\tilde{\mathbf{x}}_i$ is the corresponding textual feature (*i.e.*, privileged information), and $y_i \in \{+1, -1\}$ is the ground truth label of the instance which is unknown. Without loss of generality, we assume the positive bags are the first L^+ training bags.

In our method, we use the generalized constraints for the MIL problem [33]. As shown in [33], the relevant images usually contain a portion of positive images, while it is more likely that the irrelevant images are all negative images. Namely, we have

$$\begin{cases} \sum_{i \in \mathcal{I}_l} \frac{y_i+1}{2} \geq \sigma|\mathcal{B}_l|, & \forall Y_l = 1, \\ y_i = -1, & \forall i \in \mathcal{I}_l \text{ and } Y_l = -1, \end{cases} \tag{1}$$

where $|\mathcal{B}_l|$ is the cardinality of the bag $\mathcal{B}_l$, and $\sigma > 0$ is a predefined ratio based on prior information. In other words, each positive bag is assumed to contain at least a portion of true positive instances, and all instances in a negative bag are assumed to be negative samples.

Recall the textual descriptions associated with the training images are also noisy, so privileged information may not be always reliable as in [42,38]. Considering the labels of instances in the negative bags are known to be negative [43,33], and the results after employing noisy privileged information for the instances in the negative bags are generally worse (see our experiments in Section 4.3), we only utilize privileged information for positive bags in our method. However, it is worth mentioning that our method can be readily used to employ privileged information for the instances in all training bags.

3.2 MIL Using Privileged Information

MIL methods can be generally classified into bag-level methods [7,5] and instance-level methods [1,33]. Since bag-level methods are generally fast and effective, we focus on bag-level methods in this paper. Specifically, we take the bag-level MIL method sMIL [5] as a showcase to explain how to exploit privileged information from loosely labeled training data. We refer to our new method as *sMIL-PI*. By transforming each training bag to one training sample, the MIL problem becomes a supervised learning problem [5], because the labels of training bags are known. Such a strategy can also be applied in our sMIL-PI method.

SVM+: Before describing our sMIL-PI method, we briefly introduce the existing work SVM+. Let us denote the training data as $\{(\mathbf{x}_i, \tilde{\mathbf{x}}_i, y_i)|_{i=1}^n\}$, where $\mathbf{x}_i$ is main feature for the i-th training sample, $\tilde{\mathbf{x}}_i$ is the corresponding feature representation of privileged information which is not available for testing data, $y_i \in \{+1, -1\}$ is the class label, and n is the total number of training samples. The goal of SVM+ [42] is to learn the classifier $f(\mathbf{x}) = \mathbf{w}'\phi(\mathbf{x}) + b$, where $\phi(\cdot)$ is a nonlinear feature mapping function. Let us define another nonlinear feature mapping function $\tilde{\phi}(\cdot)$ for privileged information, and the objective of SVM+ is as follows,

$$\min_{\tilde{\mathbf{w}}, \tilde{b}, \mathbf{w}, b} \frac{1}{2}\left(\|\mathbf{w}\|^2 + \gamma\|\tilde{\mathbf{w}}\|^2\right) + C\sum_{i=1}^{n} \xi(\tilde{\mathbf{x}}_i), \tag{2}$$
$$\text{s.t. } y_i(\mathbf{w}'\phi(\mathbf{x}_i) + b) \geq 1 - \xi(\tilde{\mathbf{x}}_i), \quad \xi(\tilde{\mathbf{x}}_i) \geq 0, \quad \forall i,$$

where γ and C are the tradeoff parameters, $\xi(\tilde{\mathbf{x}}_i) = \tilde{\mathbf{w}}'\tilde{\phi}(\tilde{\mathbf{x}}_i) + \tilde{b}$ is the *slack function*, which replaces the slack variable $\xi_i \geq 0$ in the hinge loss in SVM. Such a slack function plays a role of the teacher in the training process [42]. Recall the slack variable ξ_i in SVM tells about how difficult to classify the training sample $\mathbf{x}_i$. The slack function $\xi(\mathbf{x}_i)$ is expected to model the optimal slack variable ξ_i by using privileged information analogous to the comments and explanations from the teacher in human learning [42]. Similar to SVM, SVM+ can be solved in the dual form by optimizing a quadratic programming problem.

sMIL-PI: Let us denote $\psi(\mathcal{B}_l)$ as the feature mapping function which converts a training bag into a single feature vector. The feature mapping function in sMIL is defined as the mean of instances inside the bag, *i.e.*, $\psi(\mathcal{B}_l) = \frac{1}{|\mathcal{B}_l|}\sum_{i\in\mathcal{I}_l}\phi(\mathbf{x}_i)$, where $|\mathcal{B}_l|$ is the cardinality of the bag $\mathcal{B}_l$. Recall the labels for negative instances are assumed to be negative, so we only apply the feature mapping function on the positive training bags. For ease of presentation, we denote a set of virtual training samples $\{\mathbf{z}_j|_{j=1}^m\}$, in which $\mathbf{z}_1, \ldots, \mathbf{z}_{L^+}$ are the samples mapped from the positive bags $\{\psi(\mathcal{B}_j)|_{j=1}^{L^+}\}$, the remaining samples $\mathbf{z}_{L^++1}, \ldots, \mathbf{z}_m$ are the instances $\{\phi(\mathbf{x}_i)|i \in \mathcal{I}_l, Y_l = -1\}$ in the negative bags.

When there are additional privileged information for training data, we additionally define a feature mapping function $\tilde{\psi}(\mathcal{B}_l)$ on each training bag as the mean of the instances inside the bag by using privileged information, *i.e.*, $\tilde{\mathbf{z}}_j = \tilde{\psi}(\mathcal{B}_j) = \frac{1}{|\mathcal{B}_j|}\sum_{i\in\mathcal{I}_j}\tilde{\phi}(\tilde{\mathbf{x}}_i)$ for $j = 1, \ldots, L^+$. Based on the SVM+ formulation, the objective of our sMIL-PI can be formulated as,

$$\min_{\mathbf{w},b,\tilde{\mathbf{w}},\tilde{b},\boldsymbol{\eta}} \quad \frac{1}{2}\left(\|\mathbf{w}\|^2 + \gamma\|\tilde{\mathbf{w}}\|^2\right) + C_1\sum_{j=1}^{L^+}\xi(\tilde{\mathbf{z}}_j) + C_2\sum_{j=L^++1}^{m}\eta_j, \tag{3}$$

$$\text{s.t.} \quad \mathbf{w}'\mathbf{z}_j + b \geq p_j - \xi(\tilde{\mathbf{z}}_j), \quad \forall j = 1, \ldots, L^+, \tag{4}$$

$$\mathbf{w}'\mathbf{z}_j + b \leq -1 + \eta_j, \quad \forall j = L^+ + 1, \ldots, m, \tag{5}$$

$$\xi(\tilde{\mathbf{z}}_j) \geq 0, \quad \forall j = 1, \ldots, L^+, \tag{6}$$

$$\eta_j \geq 0, \quad \forall j = L^+ + 1, \ldots, m \tag{7}$$

where $\mathbf{w}$ and b are the variables of the classifier $f(\mathbf{z}) = \mathbf{w}'\mathbf{z} + b$, γ, C_1 and C_2 are the tradeoff parameters, $\boldsymbol{\eta} = [\eta_{L^++1}, \ldots, \eta_m]'$, the slack function is defined as $\xi(\tilde{\mathbf{z}}_j) = \tilde{\mathbf{w}}'\tilde{\mathbf{z}}_j + \tilde{b}$, and p_j is the virtual label for the virtual sample $\mathbf{z}_j$. In sMIL [5], the virtual label is calculated by leveraging the instance labels of each positive bag. As sMIL assumes that there is at least one true positive sample in each positive bag, the virtual label of positive virtual sample $\mathbf{z}_j$ is $p_j = \frac{1-(|\mathcal{B}_j|-1)}{|\mathcal{B}_j|} = \frac{2-|\mathcal{B}_j|}{|\mathcal{B}_j|}$. Similarly, for our sMIL-PI using the generalized MIL constraints in (1), we can derive it as $p_j = \frac{\sigma|\mathcal{B}_j|-(1-\sigma)|\mathcal{B}_j|}{|\mathcal{B}_j|} = 2\sigma - 1$.

By introducing dual variable $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_m]'$ for the constraints in (4) and (5), and also introducing dual variable $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_{L^+}]'$ for the constraints in (6), respectively, we arrive at the dual from of (3) as follows,

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ -\mathbf{p}'\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}'(\mathbf{K}\circ\mathbf{y}\mathbf{y}')\boldsymbol{\alpha} + \frac{1}{2\gamma}(\hat{\boldsymbol{\alpha}} + \boldsymbol{\beta} - C_1\mathbf{1})'\tilde{\mathbf{K}}(\hat{\boldsymbol{\alpha}} + \boldsymbol{\beta} - C_1\mathbf{1}), \tag{8}$$

$$\text{s.t.} \quad \boldsymbol{\alpha}'\mathbf{y} = 0, \quad \mathbf{1}'(\hat{\boldsymbol{\alpha}} + \boldsymbol{\beta} - C_1\mathbf{1}) = 0, \quad \bar{\boldsymbol{\alpha}} \leq C_2\mathbf{1}, \quad \boldsymbol{\alpha} \geq \mathbf{0}, \quad \boldsymbol{\beta} \geq \mathbf{0},$$

where $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{L^+}$ and $\bar{\boldsymbol{\alpha}} \in \mathbb{R}^{m-L^+}$ are from $\boldsymbol{\alpha} = [\hat{\boldsymbol{\alpha}}', \bar{\boldsymbol{\alpha}}']'$, $\mathbf{y} = [\mathbf{1}'_{L^+}, -\mathbf{1}'_{m-L^+}]'$ is the label vector, $\mathbf{p} = [p_1, \ldots, p_{L^+}, \mathbf{1}'_{m-L^+}]' \in \mathbb{R}^m$, $\mathbf{K} \in \mathbb{R}^{m \times m}$ is the kernel matrix constructed by using the visual features, $\tilde{\mathbf{K}} \in \mathbb{R}^{L^+ \times L^+}$ is the kernel matrix constructed by using privileged information (*i.e.*, the textual features). The above problem is jointly convex in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which can be efficiently solved by optimizing a quadratic programming problem.

3.3 Domain Adaptive MIL-PI

The collected web images may have very different statistical properties with the test images (*e.g.*, the images in the Caltech-256 dataset), which is also known as the dataset bias problem [39]. To reduce domain distribution mismatch, we proposed an effective method by re-weighting the source domain samples when learning the sMIL-PI classifier. In the following, we develop our domain adaptation method, which is referred as sMIL-PI-DA.

Inspired by Kernel Mean Matching (KMM) [26], we also propose to learn the weights for the source domain samples by minimizing Maximum Mean Discrepancy (MMD) between two domains. However, KMM is a two-stage method, in which they first learn the weights for the source domain samples and then utilize the weights to train a weighted SVM. Though the recent work [8] proposed to combine the primal formulation of weighted-SVM and a regularizer based on the MMD criterion, their objective function is non-convex. Thus the global optimal solution cannot be guaranteed. To this end, we propose a convex formulation by adding the regularizer based on the MMD criterion to the dual formulation of our sMIL-PI in (8). Formally, let us denote the target domain samples as $\{\mathbf{x}_i^t|_{i=1}^{n_t}\}$, and also denote $\mathbf{z}_i^t = \phi(\mathbf{x}_i^t)$ as the corresponding nonlinear feature. To distinguish the two domains, we append a superscript s to the source domain samples, *i.e.*, $\{\mathbf{z}_i^s|_{i=1}^{m}\}$ is the set of source domain virtual samples used in our sMIL-PI-DA. We denote the objective in (8) as $H(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\mathbf{p}'\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}'(\mathbf{K} \circ \mathbf{y}\mathbf{y}')\boldsymbol{\alpha} + \frac{1}{2\gamma}(\hat{\boldsymbol{\alpha}} + \boldsymbol{\beta} - C_1\mathbf{1})'\tilde{\mathbf{K}}(\hat{\boldsymbol{\alpha}} + \boldsymbol{\beta} - C_1\mathbf{1})$ and also denote the weights for source domain samples as $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]'$. Then, we formulate our sMIL-PI-DA as follows,

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}} \; H(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \frac{\mu}{2}\|\frac{1}{m}\sum_{i=1}^{m}\theta_i\mathbf{z}_i^s - \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbf{z}_i^t\|^2 \tag{9}$$

$$\text{s.t.} \; \boldsymbol{\alpha}'\mathbf{y} = 0, \quad \mathbf{1}'(\hat{\boldsymbol{\alpha}} + \boldsymbol{\beta} - C_1\mathbf{1}) = 0, \quad \bar{\boldsymbol{\alpha}} \leq C_2\mathbf{1}, \quad \boldsymbol{\beta} \geq \mathbf{0} \tag{10}$$

$$\mathbf{0} \leq \boldsymbol{\alpha} \leq C_3\boldsymbol{\theta}, \quad \mathbf{1}'\boldsymbol{\theta} = m, \tag{11}$$

where C_3 is a parameter and θ_i is the weight for $\mathbf{z}_i^s$. The last term in (9) is a regularizer based on the MMD criterion which aims to reduce the domain distribution mismatch between two domains by reweighting the source domain samples as in KMM, and the constraints in (10) are from sMIL-PI. Note in (11), we use the box constraint $\mathbf{0} \leq \boldsymbol{\alpha} \leq C_3\boldsymbol{\theta}$ to regularize the dual variable $\boldsymbol{\alpha}$, which is similarly used in weighted SVM [26]. The second constraint $\mathbf{1}'\boldsymbol{\theta} = m$ is used to enforce the expectation of sample weights to be 1. The problem in (9) is jointly convex with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and thus we can obtain the global optimum by optimizing a quadratic programming problem.

Interestingly, the primal form of (9) is closely related to the formulation of SVM+, as described below,

Proposition 1. *The primal form of (9) is equivalent to the following problem,*

$$\min_{\mathbf{w},b,\tilde{\mathbf{w}},\tilde{b},\hat{\mathbf{w}},\hat{b},\boldsymbol{\eta}} \quad J(\mathbf{w},b,\tilde{\mathbf{w}},\tilde{b},\boldsymbol{\eta}) + \frac{\lambda}{2}\|\hat{\mathbf{w}}-\rho\mathbf{v}\|^2 + C_3\sum_{i=1}^{m}\zeta(\mathbf{z}_i^s), \tag{12}$$

$$s.t. \quad \mathbf{w}'\mathbf{z}_i^s + b \geq p_i - \xi(\tilde{\mathbf{z}}_i^s) - \zeta(\mathbf{z}_i^s), \quad \forall i = 1,\ldots,L^+, \tag{13}$$

$$\mathbf{w}'\mathbf{z}_i^s + b \leq -1 + \eta_i + \zeta(\mathbf{z}_i^s), \qquad \forall i = L^+ + 1,\ldots,m, \tag{14}$$

$$\xi(\tilde{\mathbf{z}}_i^s) \geq 0, \quad \forall i = 1,\ldots,L^+, \tag{15}$$

$$\eta_i \geq 0, \qquad \forall i = L^+ + 1,\ldots,m, \tag{16}$$

$$\zeta(\mathbf{z}_i^s) \geq 0, \quad \forall i = 1,\ldots,m, \tag{17}$$

where $J(\mathbf{w},b,\tilde{\mathbf{w}},\tilde{b},\boldsymbol{\eta}) = \frac{1}{2}\left(\|\mathbf{w}\|^2 + \gamma\|\tilde{\mathbf{w}}\|^2\right) + C_1\sum_{j=1}^{L^+}\xi(\tilde{\mathbf{z}}_j^s) + C_2\sum_{j=L^++1}^{m}\eta_j$ *is the objective function in (3),* $\zeta(\mathbf{z}_i^s) = \hat{\mathbf{w}}'\mathbf{z}_i^s + \hat{b}$, $\mathbf{v} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{z}_i^s - \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbf{z}_i^t$, $\lambda = \frac{(mC_3)^2}{\mu}$ *and* $\rho = \frac{mC_3}{\lambda}$.

Proof. We prove the dual form of (12) can be equivalently rewritten as (9). Let us introduce the dual variables $\hat{\boldsymbol{\alpha}} = [\alpha_1,\ldots,\alpha_{L^+}]' \in \mathbb{R}^{L^+}$ for the constraints in (13), $\bar{\boldsymbol{\alpha}} = [\alpha_{L^++1},\ldots,\alpha_m]' \in \mathbb{R}^{m-L^+}$ for the constraints (14), $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_{L^+}]' \in \mathbb{R}^{L^+}$ for the constraints in (15), $\boldsymbol{\tau} = [\tau_1,\ldots,\tau_{m-L^+}]' \in \mathbb{R}^{m-L^+}$ for the constraints in (16), and $\boldsymbol{\nu} = [\nu_1,\ldots,\nu_m]'$ for the constraints in (17). We also define $\boldsymbol{\alpha} = [\hat{\boldsymbol{\alpha}}',\bar{\boldsymbol{\alpha}}']'$, $\mathbf{Z} = [\mathbf{z}_1^s,\ldots,\mathbf{z}_m^s]$, $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1^s,\ldots,\tilde{\mathbf{z}}_{L^+}^s]$, and $\mathbf{y} = [\mathbf{1}_{L^+}',-\mathbf{1}_{m-L^+}']'$. By setting the derivatives of the Lagrangian of (12) w.r.t. $\mathbf{w},b,\tilde{\mathbf{w}},\tilde{b},\hat{\mathbf{w}},\hat{b},\boldsymbol{\eta}$ to zeros and substituting the derived equations back into the Lagrangian of (12), we obtain the following dual form,

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\nu}} \; -\mathbf{p}'\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}'(\mathbf{K}\circ\mathbf{y}\mathbf{y}')\boldsymbol{\alpha} + \frac{1}{2\gamma}(\hat{\boldsymbol{\alpha}}+\boldsymbol{\beta}-C_1\mathbf{1})'\tilde{\mathbf{K}}(\hat{\boldsymbol{\alpha}}+\boldsymbol{\beta}-C_1\mathbf{1}) \tag{18}$$

$$+\frac{1}{2\lambda}(\boldsymbol{\alpha}+\boldsymbol{\nu}-C_3\mathbf{1}_m)'\mathbf{K}(\boldsymbol{\alpha}+\boldsymbol{\nu}-C_3\mathbf{1}_m) + \rho\mathbf{v}'\mathbf{Z}(\boldsymbol{\alpha}+\boldsymbol{\nu}-C_3\mathbf{1}_m)$$

$$\text{s.t.} \quad \boldsymbol{\alpha}'\mathbf{y} = 0, \quad \mathbf{1}_{L^+}'(\hat{\boldsymbol{\alpha}}+\boldsymbol{\beta}-C_1\mathbf{1}_{L^+}) = 0, \quad \bar{\boldsymbol{\alpha}} \leq C_2\mathbf{1}_{m-L^+},$$

$$\mathbf{1}_m'(\boldsymbol{\alpha}+\boldsymbol{\nu}-C_3\mathbf{1}_m) = 0, \quad \boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\nu} \geq \mathbf{0}.$$

Let us define $\boldsymbol{\theta} = \frac{1}{C_3}(\boldsymbol{\alpha}+\boldsymbol{\nu})$, and the feasible set for $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\nu})$ becomes $\mathcal{A} = \{\boldsymbol{\alpha}'\mathbf{y} = 0, \mathbf{1}_{L^+}'(\hat{\boldsymbol{\alpha}}+\boldsymbol{\beta}-C_1\mathbf{1}_{L^+}) = 0, \bar{\boldsymbol{\alpha}} \leq C_2\mathbf{1}_{m-L^+}, \mathbf{1}_m'\boldsymbol{\theta} = m, \boldsymbol{\alpha} \leq C_3\boldsymbol{\theta}, \boldsymbol{\alpha},\boldsymbol{\beta} \geq \mathbf{0}\}$, then we arrive at,

$$\min_{(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{A}} -\mathbf{p}'\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}'(\mathbf{K}\circ\mathbf{y}\mathbf{y}')\boldsymbol{\alpha} + \frac{1}{2\gamma}(\hat{\boldsymbol{\alpha}}+\boldsymbol{\beta}-C_1\mathbf{1})'\tilde{\mathbf{K}}(\hat{\boldsymbol{\alpha}}+\boldsymbol{\beta}-C_1\mathbf{1}) \tag{19}$$

$$+\frac{(C_3)^2}{2\lambda}(\boldsymbol{\theta}-\mathbf{1}_m)'\mathbf{K}(\boldsymbol{\theta}-\mathbf{1}_m) + \rho C_3\mathbf{v}'\mathbf{Z}(\boldsymbol{\theta}-\mathbf{1}_m).$$

Recall that we have defined $\lambda = \frac{(C_3 m)^2}{\mu}$ and $\rho = \frac{C_3 m}{\lambda} = \frac{\mu}{C_3 m}$. By substituting the equation $\mathbf{v}'\mathbf{Z} = \frac{1}{m}\mathbf{1}'_m\mathbf{K} \; - \frac{1}{n_t}\mathbf{1}'_{n_t}\mathbf{K}_{ts}$ into the objective and replacing the constant terms with $\frac{\mu}{2n_t^2}\mathbf{1}'_{n_t}\mathbf{K}_t\mathbf{1}_{n_t}$, where $\mathbf{K}_{ts} \in \mathbb{R}^{n_t \times m}$ is the kernel matrix between the target domain samples and the source domain samples, and $\mathbf{K}_t \in \mathbb{R}^{n_t \times n_t}$ is the kernel matrix on the target domain samples, then the optimization problem in (19) finally becomes,

$$\min_{(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{A}} \; H(\boldsymbol{\alpha},\boldsymbol{\beta}) + \frac{\mu}{2}\|\frac{1}{m}\sum_{i=1}^{m}\theta_i\mathbf{z}_i^s - \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbf{z}_i^t\|^2, \tag{20}$$

where $H(\boldsymbol{\alpha},\boldsymbol{\beta})$ is defined as in (9). We complete the proof here. □

Compared with the objective function in (3), we introduce one more slack function $\zeta(\mathbf{z}_i^s) = \hat{\mathbf{w}}'\mathbf{z}_i^s + \hat{b}$, and also regularize the weight vector of this slack function by using the regularizer $\|\hat{\mathbf{w}} - \rho\mathbf{v}\|^2$. Recall that the witness function in MMD is defined as $g(\mathbf{z}) = \frac{1}{\|\mathbf{v}\|}\mathbf{v}'\mathbf{z}$ [23], which can be deemed as the mean similarity between $\mathbf{z}$ and the source domain samples (*i.e.*, $\frac{1}{m}\sum_{i=1}^{m}\mathbf{z}_i^{s\prime}\mathbf{z}$) minus the mean similarity between $\mathbf{z}$ and the target domain samples (*i.e.*, $\frac{1}{n_t}\sum_{i=1}^{n_t}\mathbf{z}_i^{t\prime}\mathbf{z}$). In other words, we conjecture that the witness function outputs a lower value when the sample $\mathbf{z}$ is closer to the target domain samples and vice versa. By using the regularizer $\|\hat{\mathbf{w}} - \rho\mathbf{v}\|^2$, we expect the new slack function $\zeta(\mathbf{z}_i^s) = \hat{\mathbf{w}}'\mathbf{z}_i^s + \hat{b}$ shares the similar trend[2] with the witness function $g(\mathbf{z}_i^s) = \frac{1}{\|\mathbf{v}\|}\mathbf{v}'\mathbf{z}_i^s$. As a result, the training error of the training sample $\mathbf{z}_i^s$ (*i.e.*, $\xi(\tilde{\mathbf{z}}_i^s) + \zeta(\mathbf{z}_i^s)$ for the samples in positive bags or $\eta_i + \zeta(\mathbf{z}_i^s)$ for negative samples) will tend to be lower if it is closer to the target domain, which is helpful for learning a more robust classifier to better predict the target domain samples.

4 Experiments

In this section, we evaluate our method sMIL-PI for image retrieval and image categorization, respectively. Then we demonstrate the effectiveness of our domain adaptation method sMIL-PI-DA for image categorization.

We extract both textual features and visual features from the training web images. The textual features are used as privileged information.

- **Textual feature:** A 200-dim term-frequency (TF) feature is extracted for each image by using the top-200 words with the highest frequency as the vocabulary. Stop-word removal is performed to remove the meaningless words.
- **Visual feature:** We extract DeCAF features [10], which has shown promising performance in various tasks. Following [10], we use the outputs from the 6th layer as visual features, which leads to 4,096-dim DeCAF$_6$ features.

In all our experiments for image retrieval and image categorization, the test data does not contain textual information. So we can only extract the same type of visual features (*i.e.*, DeCAF$_6$ features) for the images in the test set.

[2] The bias term $\hat{b}$ and the scalar terms ρ and $\frac{1}{\|\mathbf{v}\|}$ will not change the trend of functions.

4.1 Image Retrieval

Baselines: For image retrieval, we firstly compare our proposed method with two sets of baselines: the recent LUPI methods including pSVM+ [42] and Rank Transfer (RT) [38], as well as the conventional MIL method sMIL [5]. We also include SVM as a baseline, which is trained by only using the visual features. Moreover, we also compare our method with Classeme [41] and multi-view learning methods KCCA and SVM-2K, because they can also be used for our application.

- *Kernel Canonical Correlation Analysis (KCCA)* [25]: We apply KCCA on the training set by using the textual features and visual features, and then train the SVM classifier by using the common representations of visual features. In the testing process, the visual features of test samples are transformed into their common representations for the prediction.
- *SVM-2K* [16]: We train the SVM-2K classifiers by using the visual features and text features from the training samples, and apply the visual feature based classifier on the test samples for the prediction.
- *Classeme* [41]: For each word in the 200-dim textual features, we retrieve relevant and irrelevant images to construct positive bags and negative bags, respectively. Then we follow [30] to use mi-SVM to train the classeme classifier for each word. For each training image and test image, 200 decision values are obtained by using 200 learnt classeme classifiers and the decision values are augmented with the visual features. Finally, we train the SVM classifiers for classifying the test images based on the augmented features.

We also compare our method with MIML [44]. While we treat the top 200 words in the textual descriptions as noisy class labels, MIML cannot be directly applied to our task because the 200 words are not as the same as the concepts names. Thus, we use the decision values from the MIML classifiers as the features, similarly as in Classeme.

Experimental Settings. We use two web image datasets NUS-WIDE [9] and WebQuery [28] to evaluate our sMIL-PI method for image retrieval [43,33].

The NUS-WIDE dataset contains $269,648$ images, which is officially split into a training set (60%) and a test set (40%). All images in NUS-WIDE are associated with noisy tags, which are also manually annotated as 81 concepts. The WebQuery dataset contains $71,478$ web images retrieved from 353 textual queries. Each image in WebQuery is associated with textual descriptions in English or other languages (*e.g.*, French). In this work, we only use the images associated with English descriptions, and divide those images into a training set (60%) and a test set (40%). The textual queries with less than 100 training images are discarded. Finally, we obtain $19,665$ training images and $13,114$ test images from 163 remaining textual queries on the WebQuery dataset.

For both datasets, we train the classifiers using the training set and evaluate the performances of different methods on the test set. For the NUS-WIDE dataset, we follow [33] to construct 25 positive bags and 25 negative bags by respectively using relevant and irrelevant images, in which each training bag

Table 1. MAPs (%) of different methods for image retrieval. The results in boldface are from our method.

Method	Dataset	
	NUS-WIDE	WebQuery
SVM	54.41	48.51
pSVM+	57.92	50.35
RT	42.63	31.92
Classeme	54.14	48.48
MIML	54.23	48.56
KCCA	54.62	47.86
SVM-2K	54.43	49.04
sMIL	56.72	51.42
sMIL-PI	**60.88**	**52.63**

contains 15 instances. We strictly follow [33] to uniformly partition the ranked relevant images into bags. For the WebQuery dataset, we use the retrieved images from each textual query to construct the positive bags, and randomly sample the same number of images from other queries to construct the negative bags. Considering only about $100 \sim 150$ training images are retrieved from each textual query, we set the bag size as 5 to construct more training bags. Note the ground truth labels of training images are not used in the training process for both datasets. The positive ratio is set as $\sigma = 0.6$, as suggested in [33]. In our experiments, we use Gaussian kernel for visual features and linear kernel for textual features for our method and the baseline methods except RankTransfer (RT). The objective function of RT is solved in the primal form, so we can only use linear kernel instead of Gaussian kernel for visual features.

Considering the users are generally more interested in the top-ranked images, we use Average Precision (AP) based on the 100 top-ranked images for performance evaluation as suggested in [33]. The mean of APs (MAP) over all classes is used to compare different methods. We empirically fix $C_1 = C_2 = 1$ and $\gamma = 10$ for our method. For baseline methods, we choose the optimal parameters according to their MAPs on the test dataset.

Experimental Results. The MAPs of all methods are shown in Table 1. By exploiting the additional textual features, pSVM+ outperforms SVM. The multi-view learning methods KCCA and SVM-2K are also comparable or better than SVM. RankTransfer (RT) is much worse than SVM, possibly because it can only use the linear kernel. We also observe that Classeme and MIML only achieve comparable results with SVM. The sMIL method outperforms SVM, which demonstrates it is beneficial to cope with label noise by using sMIL.

Our method is better than SVM, the existing LUPI methods pSVM+ and RT, Classeme, MIML, and multi-view learning methods KCCA and SVM2K, which demonstrates the effectiveness of our sMIL-PI method for image retrieval by coping with loosely labeled web data and simultaneously taking advantage of the additional textual features as privileged information. Our sMIL-PI method also

Table 2. The left subtable lists the MAPs (%) of different methods without using domain adaptation. The right subtable reports the MAPs (%) of SVM, sMIL-PI and different domain adaptation methods. For SA, TCA, DIP, KMM, GFK and SGF, the first number is obtained by using the SVM classifiers and the second number in the parenthesis is obtained by using our sMIL-PI. The results in boldface are from our methods.

Method	Training Set	
	NUS-WIDE	Flickr
SVM	65.33	31.41
pSVM+	66.61	35.84
RT	55.53	19.09
Classeme	66.58	34.57
MIML	66.66	34.60
KCCA	65.94	35.69
SVM-2K	66.61	35.09
sMIL	67.73	35.26
sMIL-PI	**68.55**	**39.49**

Method	Training Set	
	NUS-WIDE	Flickr
SVM	65.33	31.41
sMIL-PI	68.55	39.49
sMIL-PI-DA	**70.56**	**41.35**
DASVM	67.96	33.52
STM	65.73	28.52
SA	56.13(68.73)	30.15(39.61)
TCA	61.28(66.64)	27.91(37.57)
DIP	61.08(65.32)	26.49(35.16)
KMM	60.32(68.78)	32.08(37.85)
GFK	62.98(64.60)	23.90(29.24)
SGF	66.29(68.57)	30.08(37.46)

outperforms its corresponding conventional MIL method sMIL. It again demonstrates it is beneficial to exploit the textual features as privileged information for training a more robust visual feature based classifier.

4.2 Image Categorization without Domain Adaptation

For image categorization without considering domain distribution mismatch, we use the same baselines as in image retrieval.

Experimental Settings. We evaluate our sMIL-PI method for image categorization on the benchmark dataset Caltech-256 [24]. We use the training set of NUS-WIDE as the training data. Considering different datasets contain different class names, we use their common class names for performance evaluation. Specifically, there are 17 common class names between NUS-WIDE and Caltech-256. We use the images from these 17 common classes as the test images. In total, we have $2,620$ test images for performance evaluation.

Since most of the class names in the WebQuery dataset consist of multiple words, it is ambiguous to define common classes between WebQuery and Caltech-256. So we do not use WebQuery as the training set here. Instead, we construct a new training dataset called "Flickr", in which we crawl $142,081$ Flickr images using the class names in Caltech-256 as the queries. The whole Caltech-256 dataset which contains $29,780$ images is used as the test set for performance evaluation. This setting is more challenging because we have a large number of classes and test images.

We use Average Precision (AP) based on all test images for performance evaluation. The mean of APs (MAP) over all classes is used to compare different methods. For our method, we use the same parameters as in image retrieval. For the baseline methods, we choose the optimal parameters based on their MAPs on the test dataset.

Experimental Results. The MAPs of all methods are reported in the left subtable of Table 2. As in the image retrieval application, pSVM+ is better than SVM and RT is worse than SVM. Moreover, sMIL outperforms SVM. Classeme, MIML, and Multi-view learning methods KCCA and SVM-2K are also better than SVM.

We observe that our method sMIL-PI is better than SVM, pSVM+, RT, Classeme, MIML and multi-view learning methods, which clearly demonstrates the effectiveness of our method sMIL-PI for image categorization. Moreover, our method sMIL-PI is better than its corresponding conventional MIL method sMIL, which again demonstrates it is beneficial to exploit the additional textual features as privileged information.

4.3 How to Utilize Privileged Information

As discussed in Section 3, in our sMIL-PI method, we use privileged information for relevant images (*i.e.*, positive bags) only, because privileged information (*i.e.*, textual features) may not be always reliable. To verify it, we evaluate SVM+ by utilizing privileged information for all training samples.

We report the results for image retrieval and image categorization by using NUS-WIDE as the training set. The MAPs of SVM+ and pSVM+ are 54.95% and 57.92% (*resp.*, 64.29% and 66.61%) for image retrieval (*resp.*, image categorization), which demonstrates the advantage of only utilizing privileged information for positive training bags.

4.4 Image Categorization with Domain Adaptation

Baselines. We compare our domain adaptation method sMIL-PI-DA with the existing domain adaptation methods GFK [21], SGF [22], SA [18], TCA [36], KMM [26], DIP [2], DASVM [4] and STM [8]. We notice that the feature-based domain adaptation methods such as GFK, SGF, SA, TCA, DIP can be combined with the SVM classifier or our sMIL-PI method, so we report two results by using the SVM classifier and our sMIL-PI classifier for these methods.

Experiment Settings. We use the same setting as in Section 4.2. sMIL-PI-DA has two more parameters (i.e., C_3 and λ) when compared with sMIL-PI. We empirically fix C_3 as 10 and λ as 10^4. For the baseline methods, we choose the optimal parameters based on their MAPs on the test dataset.

Experimental Results. The MAPs of all methods by using NUS-WIDE and Flickr as the training datasets are reported in the right subtable of Table 2.

The existing feature-based domain adaptation methods GFK, SGF, SA, TCA, DIP by using the SVM (*resp.*, sMIL-PI) classifier are generally comparable or even worse when compared with SVM (*resp.*, sMIL-PI). One possible explanation is the feature distributions of web images and the images from Caltech-256 are quite different. For these feature-based baselines, their results after using sMIL-PI classifier are better when compared with those using SVM classifier, which again shows the effectiveness of our sMIL-PI for image categorization by coping with label noise and simultaneously taking advantage of the additional textual features as privileged information. Moreover, DASVM is better than SVM, possibly because it can better utilize noisy training samples by progressively removing some source domain samples during the training process.

Our method is more related to KMM and STM. We also report two results for KMM because KMM can be combined with SVM or our sMIL-PI, in which the instance weights are learnt in the first step and we use the learnt instance weights to reweight the loss function of SVM or sMIL-PI in the second step. We observe that our method is better than STM and KMM with SVM or sMIL-PI, because our method can solve for the global solution while KMM is a two-step approach and STM can only achieve a local optimum.

We also observe that our method sMIL-PI-DA outperforms sMIL-PI and all the existing domain adaptation baselines, which demonstrates the advantage of our domain adaptation method sMIL-PI-DA.

5 Conclusion

In this paper, we have proposed a new method sMIL-PI for image categorization by learning from web data. Our method not only takes advantage of the additional textual features in training web data but also effectively copes with noise in the loose labels of relevant training images. We also extend sMIL-PI to handle the distribution mismatch between the training and test data, which leads to our new domain adaptation method sMIL-PI-DA. Extensive experiments for image retrieval and image categorization clearly demonstrate the effectiveness of our newly proposed methods by exploiting privileged information from web data.

Acknowledgement. This work was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by a grant from the Singapore National Research Foundation and administered by the Interactive & Digital Media Programme Office at the Media Development Authority. This work is also supported by the Singapore MoE Tier 2 Grant (ARC42/13).

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS (2003)
2. Baktashmotlagh, M., Harandi, M., Brian Lovell, M.S.: Unsupervised domain adaptation by domain invariant projection. In: ICCV (2013)
3. Bergamo, A., Torresani, L.: Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In: NIPS (2010)

4. Bruzzone, L., Marconcini, M.: Domain adaptation problems: A DASVM classification technique and a circular validation strategy. T-PAMI 32(5), 770–787 (2010)
5. Bunescu, R.C., Mooney, R.J.: Multiple instance learning for sparse positive bags. In: ICML (2007)
6. Chen, X., Shrivastava, A., Gupta, A.: NEIL: Extracting visual knowledge from web data. In: ICCV (2013)
7. Chen, Y., Bi, J., Wang, J.Z.: MILES: Multiple-instance learning via embedded instance selection. T-PAMI 28(12), 1931–1947 (2006)
8. Chu, W.S., DelaTorre, F., Cohn, J.: Selective transfer machine for personalized facial action unit detection. In: CVPR (2013)
9. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from National University of Singapore. In: CIVR (2009)
10. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A deep convolutional activation feature for generic visual recognition. In: ICML (2014)
11. Duan, L., Li, W., Tsang, I.W., Xu, D.: Improving web image search by bag-based re-ranking. T-IP 20(11), 3280–3290 (2011)
12. Duan, L., Xu, D., Tsang, I.W.: Domain adaptation from multiple sources: A domain-dependent regularization approach. T-NNLS 23(3), 504–518 (2012)
13. Duan, L., Tsang, I.W., Xu, D.: Domain transfer multiple kernel learning. T-PAMI 34(3), 465–479 (2012)
14. Duan, L., Xu, D., Tsang, I.W., Luo, J.: Visual event recognition in videos by learning from web data. T-PAMI 34(9), 1667–1680 (2012)
15. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)
16. Farquhar, J.D.R., Hardoon, D.R., Meng, H., Shawe-Taylor, J., Szedmak, S.: Two view learning: SVM-2K, theory and practice. In: NIPS (2005)
17. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from Google's image search. In: ICCV (2005)
18. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: ICCV (2013)
19. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS (2007)
20. Fouad, S., Tino, P., Raychaudhury, S., Schneider, P.: Incorporating privileged information through metric learning. T-NNLS 24(7), 1086–1098 (2013)
21. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR (2012)
22. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An unsupervised approach. In: ICCV (2011)
23. Gretton, A., KBorgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. JMLR 13, 723–773 (2012)
24. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. rep., California Institute of Technology (2007)
25. Hardoon, D.R., Szedmak, S., Shawe-taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation 16(12), 2639–2664 (2004)
26. Huang, J., Smola, A., Gretton, A., Borgwardt, K., Scholkopf, B.: Correcting sample selection bias by unlabeled data. In: NIPS (2007)
27. Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. IJCV 100(2), 134–153 (2012)
28. Krapac, J., Allan, M., Verbeek, J., Jurie, F.: Improving web image search results using query-relative classifier. In: CVPR (2010)

29. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: CVPR (2011)
30. Li, Q., Wu, J., Tu, Z.: Harvesting mid-level visual concepts from large-scale internet images. In: CVPR (2013)
31. Li, W., Duan, L., Tsang, I.W., Xu, D.: Batch mode adaptive multiple instance learning for computer vision tasks. In: CVPR, pp. 2368–2375 (2012)
32. Li, W., Duan, L., Tsang, I.W., Xu, D.: Co-labeling: A new multi-view learning approach for ambiguous problems. In: ICDM, pp. 419–428 (2012)
33. Li, W., Duan, L., Xu, D., Tsang, I.W.: Text-based image retrieval using progressive multi-instance learning. In: ICCV, pp. 2049–2055 (2011)
34. Li, W., Duan, L., Xu, D., Tsang, I.W.: Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. T-PAMI 36(6), 1134–1148 (2014)
35. Liang, L., Cai, F., Cherkassky, V.: Predictive learning with structured (grouped) data. Neural Networks 22, 766–773 (2009)
36. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. T-NN 22(2), 199–210 (2011)
37. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. T-PAMI 33(4), 754–766 (2011)
38. Sharmanska, V., Quadrianto, N., Lampert, C.H.: Learning to rank using privileged information. In: ICCV (2013)
39. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR (2011)
40. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. T-PAMI 30(11), 1958–1970 (2008)
41. Torresani, L., Szummer, M., Fitzgibbon, A.: Efficient object category recognition using classemes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 776–789. Springer, Heidelberg (2010)
42. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged infromatin. Neural Networks 22, 544–557 (2009)
43. Vijayanarasimhan, S., Grauman, K.: Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In: CVPR (2008)
44. Zhou, Z., Zhang, M.: Multi-instance multi-label learning with application to scene classification. In: NIPS (2006)