Visual Recognition by Learning from Web Data: A Weakly Supervised Domain Generalization Approach

Li Niu, Wen Li, and Dong Xu
School of Computer Engineering, Nanyang Technology University (NTU), Singapore
{lniu002,wli1,dongxu}@ntu.edu.sg

Abstract

In this work, we formulate a new weakly supervised domain generalization approach for visual recognition by using loosely labeled web images/videos as training data. Specifically, we aim to address two challenging issues when learning robust classifiers: 1) coping with noise in the labels of training web images/videos in the source domain; and 2) enhancing generalization capability of learnt classifiers to any unseen target domain. To address the first issue, we partition the training samples in each class into multiple clusters. By treating each cluster as a "bag" and the samples in each cluster as "instances", we formulate a multi-instance learning (MIL) problem by selecting a subset of training samples from each training bag and simultaneously learning the optimal classifiers based on the selected samples. To address the second issue, we assume the training web images/videos may come from multiple hidden domains with different data distributions. We then extend our MIL formulation to learn one classifier for each class and each latent domain such that multiple classifiers from each class can be effectively integrated to achieve better generalization capability. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our new approach for visual recognition by learning from web data.

1. Introduction

Recently, there is an increasing research interest in exploiting web images/videos crawled from Internet as training data to learn robust classifiers for recognizing new images/videos. However, the visual feature distributions of training and testing samples may differ considerably in terms of statistical properties, which is known as the dataset bias problem [39]. To this end, a large number of domain adaptation approaches have been proposed for various computer vision tasks [4, 44, 16, 38, 29, 22, 13, 15, 33, 14, 8, 9, 32].

Following the terminology in domain adaptation, we refer to the training dataset and testing dataset as the source domain and target domain, respectively. When target domain data is unavailable during the training process, the domain adaptation problem becomes another related task called domain generalization, which aims to learn robust classifiers that can generalize well to any unseen target domain [27, 36, 43]. Domain generalization is also an important research problem for the real-world visual recognition applications. For example, the datasets containing the photos/videos from each user can be considered as one target domain because different users may use different cameras to capture the photos/videos in their own ways. So we have a large number of target domains from various users and meanwhile some users may not be willing to share their photos/videos to others as target domain data due to the privacy issue. In this case, it is more desirable to develop new domain generalization approaches without using target domain data in the training process.

In this work, we study the domain generalization problem by exploiting web images/videos as source domain data. In Section 3, we propose an effective weakly supervised domain generalization (WSDG) approach to address this problem. In our approach, we consider two important issues when exploiting web images/videos as source domain data: 1) the training web images and videos are often associated with inaccurate labels, so the learnt classifiers may be less robust, and the recognition performance may be significantly degraded as well; 2) the test data in the target domain usually has a different distribution from the training images/videos, and the target domain data is often unavailable in the training stage.

Specifically, to cope with label noise of web training images and videos, we partition the training samples from each class into a set of clusters, and then treat each cluster as a "bag" and the samples in each cluster as "instances". Inspired by the multi-instance learning (MIL) methods, we only know the labels of training bags, while the labels of instances within each training bag remain unknown. Then, we aim to select a subset of training samples from each training bag to represent this bag, such that the training bags from all

the classes can be well separated. To this end, we propose a new multi-class MIL formulation to learn the classifiers from multiple classes and select the training samples from each training bag.

On the other hand, we assume the training web images/videos may come from multiple hidden domains with distinctive data distributions, as suggested in the recent works [24, 20, 43]. After multiple latent domains are discovered with the existing technology (*e.g.*, [20]), we aim to learn one classifier for each class and each latent domain. As each classifier is learnt from the training samples with a distinctive data distribution, each integrated classifier, which is obtained by combining multiple classifiers from each class, is expected to be robust to the variation of data distributions, and thus can be well generalized to predict test data from any unseen target domain. Recall that we only use a subset of training samples from each training bag for learning the classifiers, we further introduce a regularizer based on the maximum mean discrepancy (MMD) criterion to select the training samples with more distinctive data distributions in order to further enhance domain generalization ability.

Moreover, the web images and videos are generally associated with valuable contextual information (*e.g.*, tags, captions, and surrounding texts). Although such textual descriptions are usually not available in the testing images and videos, they can still be used as privileged information [40, 33]. We further extend our WSDG approach by taking advantage of the additional textual descriptions as privileged information. In Section 4, we conduct comprehensive experiments for visual recognition by learning from web data, and the results clearly demonstrate the effectiveness of our newly proposed approaches.

2. Related Work

Our work is related to the multi-instance learning methods in [41, 31, 30, 33], which explicitly coped with noise in the loose labels of web images/videos. Particularly, the training images/videos are partitioned into a set of clusters and each cluster is treated as a "bag" with the images/videos in each bag as "instances". As a result, this task can be formulated as a MIL problem and different MIL methods were proposed in [41, 31, 30]. Other popular MIL methods include mi-SVM [2], which uses a heuristic way to iteratively train the SVM classifier and infer the instance labels, and KI-SVM [34], which aims to discover the key instance inside each bag to represent the bag. However, these works did not consider the domain generalization task when learning the SVM classifiers, so they may not generalize well to any unseen target domain.

Our work is also related to domain generalization. For domain generalization, Muandet *et al*. [36] proposed to learn domain invariant feature representations, while Khosla *et al*. [27] proposed an SVM based approach. Xu *et al*. [43] proposed an exemplar SVM based method by exploiting the low-rank structure in the source domain. When target domain data is available in the training phase, domain adaptation methods were recently developed to reduce the domain distribution mismatch, and these methods can be roughly categorized into feature-based methods [29, 22, 21, 3, 18], classifier-based methods [5, 15, 13], and instance-reweighting methods [25]. Interested readers can refer to the recent survey [37] for more details.

Our work is more related to the recent works for discovering latent domains [20, 24, 43]. The source domain is divided into different latent domains by using a clustering based approach in [24] or the MMD criterion in [20]. After the latent domains are discovered, an NN or SVM classifier is trained for each latent domain separately. Finally, all classifiers are fused to predict the target domain samples. In contrast, we jointly learn multiple classifiers which are effectively integrated for each class to maximize the separability between different classes, leading to better generalization performances.

Moreover, our work is related to the sub-categorization problem [23], which assumes several subcategories exist in each class. Some researchers also attempt to combine sub-categorization with multi-instance learning [42, 45, 46]. However, these works did not consider the domain distribution mismatch problem between the training and test data. In contrast, the domain generalization problem aims to handle the test data from different domains with large distribution variations. So their motivations and formulations are intrinsically different from our work.

3. Weakly Supervised Domain Generalization

In this section, we propose our weakly supervised domain generalization (WSDG) method, in which we simultaneously learn robust classifiers and select good samples by removing the outliers (*i.e.*, training samples with inaccurate class labels). For ease of presentation, a vector/matrix is denoted by a lowercase/uppercase letter in boldface. The transpose of a vector/matrix is denoted using the superscript $'$. We also denote $\mathbf{0}_n, \mathbf{1}_n \in \mathbb{R}^n$ as the n-dim column vectors of all zeros and all ones, respectively. When the dimensionality is obvious, we use $\mathbf{0}$ and $\mathbf{1}$ instead of $\mathbf{0}_n$ and $\mathbf{1}_n$ for simplicity. Moreover, we denote $\mathbf{A} \circ \mathbf{B}$ as the element-wise product between two matrices. The inequality $\mathbf{a} \leq \mathbf{b}$ means that $a_i \leq b_i$ for $i = 1, \ldots, n$. The indicator function is represented as $\delta(a = b)$, where $\delta(a = b) = 1$ if $a = b$, and $\delta(a = b) = 0$, otherwise.

Suppose the source domain data contains N training samples belonging to C classes, we denote the source domain data as $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i$ is the i-th training sample, and $y_i \in \{1, \ldots, C\}$ is the corresponding class label of $\mathbf{x}_i$. In the following, we firstly introduce

how to discover latent domains by using the existing technology [20], and then propose a multi-class multi-instance learning method without considering the latent domain issues. Finally, we incorporate the discovered latent domains into our multi-class multi-instance learning approach in order to learn the integrated classifiers, which are robust to any unseen target domain.

3.1. Discovering Latent Domains

In this work, we employ the latent domain discovering method [20], which is based on the Maximum Mean Discrepancy (MMD) criterion. We denote $\pi_{i,m} \in \{0,1\}$ as the indicator, namely, $\pi_{i,m} = 1$ if $\mathbf{x}_i$ belongs to the m-th latent domain, and $\pi_{i,m} = 0$ otherwise. Let us define $N_m = \sum_{i=1}^{N} \pi_{i,m}$ as the number of samples in the m-th latent domain. In [20], the goal is to make the discovered latent domains as distinctive as possible, which can be formulated as the following optimization problem by maximizing the sum of MMDs between any two latent domains,

$$\max_{\pi_{i,m}} \sum_{m \neq \tilde{m}} \| \frac{1}{N_m} \sum_{i=1}^{N} \pi_{i,m} \phi(\mathbf{x}_i) - \frac{1}{N_{\tilde{m}}} \sum_{i=1}^{N} \pi_{i,\tilde{m}} \phi(\mathbf{x}_i) \|^2, \quad (1)$$

where $\phi(\cdot)$ is the feature mapping function induced by a kernel $\mathbf{K} \in \mathbb{R}^{N \times N}$ on the training samples (*i.e.*, $\mathbf{K} = [K_{i,j}]$ where $K_{i,j} = \phi(\mathbf{x}_i)'\phi(\mathbf{x}_j)$). Let $\beta_{i,m} = \frac{\pi_{i,m}}{N_m}$ and $\boldsymbol{\beta}_m = [\beta_{1,m}, \ldots, \beta_{N,m}]'$, the above problem can be relaxed as the following optimization problem [20],

$$\max_{\boldsymbol{\beta}} \quad \sum_{m \neq \tilde{m}} (\boldsymbol{\beta}_m - \boldsymbol{\beta}_{\tilde{m}})' \mathbf{K} (\boldsymbol{\beta}_m - \boldsymbol{\beta}_{\tilde{m}}) \quad (2)$$

$$\text{s.t.} \quad \frac{1}{N} \leq \sum\nolimits_{m=1}^{M} \beta_{i,m} \leq \frac{1}{C}, \quad \forall i, \quad (3)$$

$$\sum\nolimits_{i=1}^{N} \delta(y_i = c) \beta_{i,m} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \delta(y_i = c), \forall c, m, \quad (4)$$

$$\sum\nolimits_{i=1}^{N} \beta_{i,m} = 1, \beta_{i,m} \geq 0, \quad \forall i, m, \quad (5)$$

where the first constraint in (3) is to enforce each latent domain to contain at least one training sample per class, the second constraint in (4) is to keep the sample proportion of each class in each latent domain the same as that in the whole source domain, and the third constraint in (5) is from the definitions of $\beta_{i,m}$ and $\pi_{i,m}$. We refer interested readers to [20] for the detailed derivations. Although the above problem is a non-convex quadratic programming problem, it can still be solved by using the existing solvers to obtain a satisfactory solution [20, 1].

After discovering latent domains by solving the problem in (2), we then learn one classifier for each class and each latent domain, and integrate multiple classifiers from each class based on the learnt $\beta_{i,m}$'s. In the following section, we firstly discuss how to cope with the label noise problem by proposing a new multi-class multi-instance learning (MIL) formulation, and then discuss how to extend our multi-class MIL formulation to enhance domain generalization ability of learnt classifiers to any unseen target domain.

3.2. Formulation

Recall that in our task, the labels of training data are noisy, so we employ the multi-instance learning (MIL) method, which only requires weakly supervised information when training the classifiers. Considering the latent domain discovery method utilizes the whole training data for discovering latent domains and includes the class balance constraint in (4), we first formulate an effective multi-class MIL method, and then propose a unified formulation to learn robust classifiers by coping with label noise and taking advantage of multiple latent domains.

3.2.1 Learning with Weakly Supervised Information

In MIL, training data is organized as a set of bags of training instances. We only have the labels of training bags, while the labels of training instances are assumed to be unknown. We partition our training samples in each class into different clusters and treat each cluster as a training bag, *i.e.*, $\{(\mathcal{B}_l, Y_l) | l = 1, \ldots, L\}$. In this work, the training images/videos are collected from the Internet by using tag-based image/video search, so the bag label $Y_l \in \{1, \ldots, C\}$ is determined by using the corresponding query tag. Usually, we assume there are at least a portion of true positive instances in each positive bag [31]. So we also define the ratio η to represent the proportion of training instances in each training bag, in which their instance labels are consistent with the bag-level label Y_l. Similarly as in MIL, η can be estimated according to the prior knowledge.

To effectively learn robust classifiers, we propose to select good samples from each bag to learn robust classifiers by removing the outliers in the training data. Let us use a binary indicator $h_i \in \{0,1\}$ to indicate whether the training sample $\mathbf{x}_i$ is selected or not. Namely, $h_i = 1$ if $\mathbf{x}_i$ is selected, and $h_i = 0$, otherwise. For ease of presentation, we denote $\mathbf{h} = [h_1, \ldots, h_N]'$ as the indicator vector. Consequently, the feasible set of the indicator vector $\mathbf{h}$ can be represented as $\mathcal{H} = \{\mathbf{h} | \sum_{i \in \mathcal{I}_l} h_i = \eta |\mathcal{B}_l|, \forall l\}$, where $\mathcal{I}_l$ is the set of indices of instances in the bag $\mathcal{B}_l$, and $|\mathcal{B}_l|$ denotes the number of instances in the training bag $\mathcal{B}_l$.

We formulate our multi-class MIL problem based on multi-class SVM [11]. Specifically, we propose to learn C classifiers $\{f_c(\mathbf{x}) | c = 1, \ldots C\}$, where each $f_c(\mathbf{x})$ is the classifier for the c-th class. We represent each classifier [1] as

[1] The bias term is omitted here for better representation. In our work, we augment the feature vector of each training sample with an additional entry of 1.

$f_c(\mathbf{x}) = (\mathbf{w}_c)'\phi(\mathbf{x})$. Inspired by the multi-class SVM [11] and the multi-instance learning method KI-SVM [34], we present a unified formulation to jointly learn $\mathbf{h}$ and C classifiers as follows,

$$\min_{\substack{\mathbf{h}\in\mathcal{H}\\ \mathbf{w}_c,\xi_l}} \quad \frac{1}{2}\sum_{c=1}^{C}\|\mathbf{w}_c\|^2 + C_1\sum_{l=1}^{L}\xi_l \tag{6}$$

$$\text{s.t.} \quad \frac{1}{|\mathcal{B}_l|}\sum_{i\in I_l} h_i\left((\mathbf{w}_{Y_l})'\phi(\mathbf{x}_i) - (\mathbf{w}_{\tilde{c}})'\phi(\mathbf{x}_i)\right)$$

$$\geq \eta - \xi_l, \quad \forall l, \tilde{c} \neq Y_l, \tag{7}$$

$$\xi_l \geq 0, \qquad \forall l, \tag{8}$$

where C_1 is a tradeoff parameter, ξ_l's are slack variables. The constraint in (7) is to encourage the total decision value of each bag $\mathcal{B}_l$ from the classifier corresponding to its ground-truth label should be greater than that from the classifier of any other class. Intuitively, good instances (*i.e.*, non-outliers) tend to contribute smaller bag-level loss (*i.e.*, ξ_l), while the outliers tend to contribute larger bag-level loss. Therefore, we learn the indicator vector $\mathbf{h}$ to select good instances from each bag in order to minimize the total bag-level loss.

Note the problem in (6) reduces to the multi-class SVM [11] when setting the bag size $|\mathcal{B}_l|$ as 1 and setting each indicator $h_i = 1$. Moreover, for the binary class classification setting, the problem in (6) can reduce to the multi-instance learning method KI-SVM [34] after minor modifications.

3.2.2 Weakly Supervised Domain Generalization

When the source domain consists of M latent domains, it is more desirable to learn one classifier for each class and each latent domain to enhance generalization ability of learnt classifiers to the test data from any unseen target domain (see the discussion in Section 3.2.3).

In particular, we propose to learn $C \times M$ classifiers $\{f_{c,m}(\mathbf{x})|c = 1,\dots C, \text{ and } m = 1,\dots,M\}$, where the classifier for the c-th class and the m-th latent domain is represented $f_{c,m}(\mathbf{x}) = (\mathbf{w}_{c,m})'\phi(\mathbf{x})$. Then the decision function on the training sample $\mathbf{x}_i$ for each class can be obtained by integrating the classifiers from different domains as $f_c(\mathbf{x}_i) = \sum_{m=1}^{M}\hat{\beta}_{i,m}f_{c,m}(\mathbf{x}_i)$, where $\hat{\beta}_{i,m}$ is the probability that the i-th training sample belongs to the m-th latent domain. Such probabilities can be precomputed by using the latent domain discover method in [20]. Specifically, we can calculate each corresponding $\hat{\beta}_{i,m}$ as $\hat{\beta}_{i,m} = \frac{\beta_{i,m}}{\sum_{m=1}^{M}\beta_{i,m}}$, where $\beta_{i,m}$'s are obtained by solving the optimization problem in (2). Our goal is to learn those $C \times M$ classifiers such that the discriminability of the integrated classifiers $f_c(\mathbf{x}_i)$'s is maximized by only using weakly labeled training data.

Moreover, one problem when using the latent domain discovery method in [20] is that the objective function in (2) is originally designed for training data with clean labels. When there are outliers in training data, it is more desirable to seek for an optimal $\mathbf{h}$ to remove the outliers such that the objective in (2) can be maximized. For ease of presentation, let us denote $\mathbf{B} = [\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_M] \in \mathbb{R}^{N\times M}$, then the objective in (2) can be written as $\rho(\mathbf{B},\mathbf{K}) = \sum_{m\neq\tilde{m}}(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{\tilde{m}})'\mathbf{K}(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{\tilde{m}})$. To learn an optimal indicator $\mathbf{h}$, we introduce a new regularizer $\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}\mathbf{h}'))$. Then we arrive at the final formulation of our WSDG method for learning the optimal classifiers $f_{c,m}(\mathbf{x})$'s as follows,

$$\min_{\substack{\mathbf{h}\in\mathcal{H}\\ \mathbf{w}_{c,m},\xi_l}} \quad \frac{1}{2}\sum_{c=1}^{C}\sum_{m=1}^{M}\|\mathbf{w}_{c,m}\|^2 + C_1\sum_{l=1}^{L}\xi_l$$

$$-C_2\,\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}\mathbf{h}')) \tag{9}$$

$$\text{s.t.} \quad \frac{1}{|\mathcal{B}_l|}\sum_{i\in I_l} h_i\left(\sum_{m=1}^{M}\hat{\beta}_{i,m}(\mathbf{w}_{Y_l,m})'\phi(\mathbf{x}_i) - (\mathbf{w}_{\tilde{c},\tilde{m}})'\phi(\mathbf{x}_i)\right)$$

$$\geq \eta - \xi_l, \qquad \forall l,\tilde{m},\tilde{c} \neq Y_l, \tag{10}$$

$$\xi_l \geq 0, \qquad \forall l. \tag{11}$$

where C_2 is a tradeoff parameter. The constraint in (10) can be explained similarly as that in (6) except that we replace $(\mathbf{w}_{Y_l})'\phi(\mathbf{x}_i)$ in (6) with $\sum_{m=1}^{M}\hat{\beta}_{i,m}(\mathbf{w}_{Y_l,m})'\phi(\mathbf{x}_i)$ and $(\mathbf{w}_{\tilde{c}})'\phi(\mathbf{x}_i)$ with $(\mathbf{w}_{\tilde{c},\tilde{m}})'\phi(\mathbf{x}_i)$.

3.2.3 Discussion

Why Latent Domain Works: In order to learn robust classifiers that can generalize well to any unseen target domain, we train one classifier for each latent domain and each class after softly partitioning the training samples into different latent domains by using the existing latent domain discover technology in [20]. Since the training samples from the same class and the same latent domain are usually with more similar data distribution [20], it is easier to learn a discriminative classifier for each class and each latent domain. In the testing process, we predict the label of a given test sample $\mathbf{x}$ by using $y = arg\max_c \max_m (\mathbf{w}_{c,m})'\phi(\mathbf{x})$. Namely, for each class, we use the best classifier with the highest decision value among the classifiers from all latent domains in order to find the best matched latent domain for this test sample. By finding the best matched latent source domain for each test sample, we conjecture the classifiers learnt by using our WSDG method can generalize well to test data from any unseen target domain.

Utilizing Privileged Information: The web images and videos are often associated with rich and valuable contextual information (*e.g.*, tags, captions, and surrounding texts). Such textual descriptions which are not available in the testing images and videos can still be used as privileged information for improving the learnt classifiers in the training

phase [40, 33]. To this end, we extend our WSDG approach to WSDG-PI by additionally using the textual features extracted from the contextual descriptions of web images and videos as privileged information. Specifically, inspired by [40, 33], we replace the loss ξ_l in (10) and (11) with $\frac{1}{|\mathcal{B}_l|}\sum_{i\in I_l} h_i(\sum_{m=1}^{M}\hat{\beta}_{i,m}(\tilde{f}_{Y_l,m}(\mathbf{z}_i)-\tilde{f}_{\tilde{c},\tilde{m}}(\mathbf{z}_i))$, where $\mathbf{z}_i$ is the textual features for the i-th training sample and the slack function $\tilde{f}_{c,m}(\mathbf{z}) = (\tilde{\mathbf{w}}_{c,m})'\phi(\mathbf{z})$ is similarly defined as $f_{c,m}(\mathbf{x})$ in (9). An additional regularizer term $\sum_{c=1}^{C}\sum_{m=1}^{M}\|\tilde{\mathbf{w}}_{c,m}\|^2$ is also added in the the objective to control the model complexity. Note the solution to the optimization problem in (9), which will be discussed below, can be similarly used to solve the optimization problem of our WSDG-PI approach.

3.3. Optimization

The problem in (9) is a nontrivial non-convex mixed integer optimization problem. Inspired by the recent works on MIL [31][34], we relax the dual form of (9) as a multiple kernel learning (MKL) problem, which can be solved similarly as in [28]. In the following, we first present the relaxation in the dual form of (9), and then introduce the detailed solution to the relaxed problem.

3.3.1 Reformulation in the Dual Form

Proposition 1. *The dual form of (9) can be written as follows,*

$$\min_{\mathbf{h}\in\mathcal{H}}\max_{\boldsymbol{\alpha}} -\frac{1}{2}\boldsymbol{\alpha}'\mathbf{Q}^{\mathbf{h}}\boldsymbol{\alpha}+\boldsymbol{\zeta}'\boldsymbol{\alpha}-C_2\,\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}\mathbf{h}')) \quad (12)$$
$$s.t. \quad \sum_{c,m}\alpha_{l,c,m} = C_1, \quad \forall l$$
$$\alpha_{l,c,m}\geq 0, \quad \forall l,c,m,$$

where $\boldsymbol{\alpha}\in\mathbb{R}^{\tilde{D}}$ is a vector with each entry being the dual variable $\alpha_{l,c,m}$, $\tilde{D} = L\cdot C\cdot M$, $\boldsymbol{\zeta}\in\mathbb{R}^{\tilde{D}}$ is a vector with each entry $\zeta_{l,c,m} = \eta$ if $c\neq Y_l$ and $\zeta_{l,c,m} = 0$ otherwise, and $\mathbf{Q}^{\mathbf{h}}\in\mathbb{R}^{\tilde{D}\times\tilde{D}}$ is a matrix in which each element can be obtained by $Q^{\mathbf{h}}_{u,v} = \frac{1}{|\mathcal{B}_l||\mathcal{B}_{\tilde{l}}|}\sum_{i\in\mathcal{I}_l}\sum_{j\in\mathcal{I}_{\tilde{l}}} h_i h_j\phi(\mathbf{x}_i)'\phi(\mathbf{x}_j)\gamma(i,j,c,\tilde{c},m,\tilde{m})$, $u = (l-1)\cdot C\cdot M+(c-1)\cdot M+m$, $v = (\tilde{l}-1)\cdot C\cdot M+(\tilde{c}-1)\cdot M+\tilde{m}$, and $\gamma(i,j,c,\tilde{c},m,\tilde{m})$ is a scalar which can be calculated as $\gamma(i,j,c,\tilde{c},m,\tilde{m}) = [1-\delta(c=y_i)][1-\delta(\tilde{c}=y_j)][\delta(y_i=y_j)\sum_{q=1}^{M}\hat{\beta}_{i,q}\hat{\beta}_{j,q}+\delta(c=\tilde{c})\delta(m=\tilde{m})]-[1-\delta(c=\tilde{c})]\{[1-\delta(c=y_i)]\delta(c=y_j)\hat{\beta}_{j,m}+[1-\delta(\tilde{c}=y_j)]\delta(\tilde{c}=y_i)\hat{\beta}_{i,\tilde{m}}\}$.

Proof. In order to simplify the primal form (9), we firstly introduce an intermediate variable $\theta_{i,c,m,\tilde{m}}$ as follows,

$$\theta_{i,c,m,\tilde{m}} = \begin{cases}\hat{\beta}_{i,m} & c = y_i,\\ \delta(m=\tilde{m}) & c\neq y_i.\end{cases} \quad (13)$$

Note here $y_i = Y_l$, $\forall i\in\mathcal{I}_l$. Then, we have $\sum_{m=1}^{M}\hat{\beta}_{i,m}(\mathbf{w}_{y_i,m})'\phi(\mathbf{x}_i) = \sum_{m=1}^{M}\theta_{i,y_i,m,\tilde{m}}(\mathbf{w}_{y_i,m})'\phi(\mathbf{x}_i)$ and $(\mathbf{w}_{c,\tilde{m}})'\phi(\mathbf{x}_i) = \sum_{m=1}^{M}\theta_{i,c,m,\tilde{m}}(\mathbf{w}_{c,m})'\phi(\mathbf{x}_i)$. Then, the constraints (10) and (11) can be uniformly written as

$$\frac{1}{|B_l|}\sum_{i\in I_l} h_i\Big(\sum_{m=1}^{M}\theta_{i,Y_l,m,\tilde{m}}(\mathbf{w}_{Y_l,m})'\phi(\mathbf{x}_i) - \sum_{m=1}^{M}\theta_{i,c,m,\tilde{m}}(\mathbf{w}_{c,m})'\phi(\mathbf{x}_i)\Big) \geq \zeta_{l,c,\tilde{m}}-\xi_l, \quad \forall l,c,\tilde{m}, \quad (14)$$

in which $\zeta_{l,c,m}$ is defined below (12).

We concatenate all $\mathbf{w}_{c,m}$'s and define $\mathbf{w} = [\mathbf{w}'_{1,1},\ldots,\mathbf{w}'_{1,M},\mathbf{w}'_{2,1},\ldots,\mathbf{w}'_{C,M}]'$. We also define a new mapping function for each $\mathcal{B}_l$ as $\varphi(\mathbf{h},\mathcal{B}_l,c,\tilde{m}) = [\frac{1}{|B_l|}\sum_{i\in I_l} h_i\,\theta_{i,1,1,\tilde{m}}\delta(c=1)\phi(\mathbf{x}_i)',\ldots,\frac{1}{|B_l|}\sum_{i\in I_l} h_i\,\theta_{i,C,M,\tilde{m}}\delta(c=C)\phi(\mathbf{x}_i)']'$. By further denoting $\psi(\mathbf{h},\mathcal{B}_l,c,\tilde{m}) = \varphi(\mathbf{h},\mathcal{B}_l,Y_l,\tilde{m})-\varphi(\mathbf{h},\mathcal{B}_l,c,\tilde{m})$, the problem in (9) can be simplified as,

$$\min_{\substack{\mathbf{h}\in\mathcal{H}\\ \mathbf{w},\xi_l}} \frac{1}{2}\|\mathbf{w}\|^2+C_1\sum_{l=1}^{L}\xi_l - C_2\,\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}\mathbf{h}')) \quad (15)$$
$$\text{s.t. } \mathbf{w}'\psi(\mathbf{h},\mathcal{B}_l,c,m)\geq\zeta_{l,c,m}-\xi_l, \quad \forall l,c,m. \quad (16)$$

Let us introduce a dual variable $\alpha_{l,c,m}$ for each constraint in (16), then we arrive at its Lagrangian as follows,

$$\mathcal{L}_{\mathbf{w},\xi_l,\alpha_{l,c,m}} = \frac{1}{2}\|\mathbf{w}\|^2+C_1\sum_{i=1}^{N}\xi_i - C_2\,\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}\mathbf{h}')) - \sum_{l,c,m}\alpha_{l,c,m}\left(\mathbf{w}'\psi(\mathbf{h},\mathcal{B}_l,c,m)-\zeta_{l,c,m}+\xi_l\right) \quad (17)$$

By setting the derivatives of $\mathcal{L}$ w.r.t. $\mathbf{w}$ and ξ_l as zeros, and substituting the obtained equalities back into (17), we can arrive at the dual form (12). □

The problem in (12) is a mixed integer programming problem, which is NP hard. So it is difficult to solve it. Inspired by the recent works on multi-instance learning, instead of directly optimizing over the indicator vector $\mathbf{h}$, we alternatively seek for an optimal linear combination of $\mathbf{h}_t\mathbf{h}_t'$'s based on all feasible $\mathbf{h}_t\in\mathcal{H}$, *i.e.*, $\sum_{\mathbf{h}_t\in\mathcal{H}} d_t\mathbf{h}_t\mathbf{h}_t'$, where d_t is the combination coefficient for $\mathbf{h}_t\mathbf{h}_t'$. For ease of presentation, we denote $T = |\mathcal{H}|$, $\mathbf{d} = [d_1,\ldots,d_T]'$, and $\mathcal{D} = \{\mathbf{d}|\mathbf{d}'\mathbf{1}=1,\mathbf{d}\geq 0\}$ as the feasible set of $\mathbf{d}$. We also

denote $\mathcal{A}$ as the feasible set of $\boldsymbol{\alpha}$ in (12). Then, we obtain the following optimization problem,

$$\min_{\mathbf{d}\in\mathcal{D}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\sum_{t=1}^{T} d_t\boldsymbol{\alpha}'\mathbf{Q}^{\mathbf{h}_t}\boldsymbol{\alpha}+\boldsymbol{\zeta}'\boldsymbol{\alpha} -C_2\sum_{t=1}^{T} d_t\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}_t\mathbf{h}_t')), \quad (18)$$

where we move the sum operator over d_t outside $\mathbf{Q}^{\mathbf{h}_t}$ and $\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}_t\mathbf{h}_t'))$, because both are linear terms of $\mathbf{h}_t\mathbf{h}_t'$. The above problem shares a similar form with the dual of the MKL problem, where each base kernel is $\mathbf{Q}^{\mathbf{h}_t}$. We therefore solve it by optimizing a convex optimization problem in its primal form as follows,

$$\min_{\mathbf{d}\in\mathcal{D},\mathbf{w}_t,\xi_l} \frac{1}{2}\sum_{t=1}^{T}\frac{\|\mathbf{w}_t\|^2}{d_t}+C_1\sum_{l=1}^{L}\xi_l -C_2\sum_{t=1}^{T} d_t\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}_t\mathbf{h}_t')) \quad (19)$$

$$\text{s.t.}\sum_{t=1}^{T}\mathbf{w}_t'\psi(\mathbf{h}_t,\mathcal{B}_l,c,m)\geq\zeta_{l,c,m}-\xi_l,\forall l,c,m, \quad (20)$$

where $\psi(\mathbf{h}_t,\mathcal{B}_l,c,m)$ is defined above (15), and $\mathbf{w}_t$ can be obtained by

$$\mathbf{w}_t = d_t\sum_{l,c,m}\alpha_{l,c,m}\psi(\mathbf{h}_t,\mathcal{B}_l,c,m). \quad (21)$$

3.3.2 The Solution to (19)

As the problem in (19) is a convex problem, we solve it by alternatively updating $\mathbf{d}$ and $\mathbf{w}_t$.

Update $\mathbf{d}$: To solve $\mathbf{d}$, we firstly write the Lagrangian of (19) by introducing a dual variable τ for the constraint $\mathbf{d}'\mathbf{1}=1$. By setting the derivative of the Lagrangian w.r.t. each d_t to zero, we obtain

$$\tau=\frac{\|\mathbf{w}_t\|^2}{2d_t^2}+C_2\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}_t\mathbf{h}_t')),\ \forall t=1,\ldots,T. \quad (22)$$

We rewrite (22) as follows,

$$d_t=\frac{\|\mathbf{w}_t\|}{\sqrt{2\tau-2C_2\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}_t\mathbf{h}_t'))}},\ \forall t=1,\ldots,T. \quad (23)$$

Note that the righthand side of (23) is a monotonically decreasing function w.r.t. τ. Considering the constraint $\mathbf{d}'\mathbf{1}=1$, we use binary search to find the value τ such that $\sum_{t=1}^{T} d_t = 1$ is satisfied. Then we can calculate d_t's by using (23).

Update $\mathbf{w}_t$: When fixing $\mathbf{d}$, we solve $\boldsymbol{\alpha}$ in its dual form (18) and recover $\mathbf{w}_t$ by using (21). The problem in (18) is a quadratic programming problem w.r.t. $\boldsymbol{\alpha}$, which can be solved by using the existing QP solvers. However, the existing QP solvers are very inefficient due to too many variables (*i.e.*, $L\cdot C\cdot M$ variables). Inspired by [7] and [17], we develop a sequential minimal optimization (SMO) algorithm to solve this QP problem.

3.3.3 Cutting-Plane Algorithm

The major challenge when using the above alternating optimization procedure is that there are too many base kernels. Inspired by Infinite Kernel Learning (IKL) [19], we employ the cutting-plane algorithm, in which we start from a small set of base kernels, and at each iteration we iteratively add a new kernel that violates the constraints. As we only need to solve an MKL problem based on a small set of $\mathbf{h}$ at each iteration, the optimization procedure is much more efficient. In particular, let us introduce a dual variable τ for the constraint $\mathbf{d}'\mathbf{1}=1$ and replace $\rho(\mathbf{B},\mathbf{K}\circ(\mathbf{h}_t\mathbf{h}_t'))$ equivalently with $\mathbf{h}_t'\mathbf{P}\mathbf{h}_t$ by defining $\mathbf{P}=\sum_{m\neq\tilde{m}}\mathbf{K}\circ((\boldsymbol{\beta}_m-\boldsymbol{\beta}_{\tilde{m}})(\boldsymbol{\beta}_m-\boldsymbol{\beta}_{\tilde{m}})')$. Then we can rewrite (18) to its equivalent form as follows,

$$\max_{\tau,\boldsymbol{\alpha}\in\mathcal{A}} -\tau+\boldsymbol{\zeta}'\boldsymbol{\alpha}, \quad (24)$$

$$\text{s.t.}\ \frac{1}{2}\boldsymbol{\alpha}'\mathbf{Q}^{\mathbf{h}_t}\boldsymbol{\alpha}+C_2\mathbf{h}_t'\mathbf{P}\mathbf{h}_t\leq\tau,\ \forall t, \quad (25)$$

which has many constraints.

To solve (24), we start to solve $\boldsymbol{\alpha}$ by using only one constraint, and then iteratively add a new constraint which is violated by the current $\boldsymbol{\alpha}$. In particular, as each constraint is associated with an $\mathbf{h}_t$, so we can find the violated constraint by optimizing the following problem,

$$\max_{\mathbf{h}}\frac{1}{2}\boldsymbol{\alpha}'\mathbf{Q}^{\mathbf{h}}\boldsymbol{\alpha}+C_2\mathbf{h}'\mathbf{P}\mathbf{h} \quad (26)$$

After a simple deduction, we can rewrite (26) as follows,

$$\max_{\mathbf{h}}\mathbf{h}'(\frac{1}{2}\hat{\mathbf{Q}}\circ(\hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}')+C_2\mathbf{P})\mathbf{h}, \quad (27)$$

where $\hat{\mathbf{Q}}\in\mathbb{R}^{N\times N}$ is the shrinked $\mathbf{Q}$ with its element $\hat{Q}_{i,j}=\sum_{c,\tilde{c},m,\tilde{m}}\gamma(i,j,c,\tilde{c},m,\tilde{m})\phi(\mathbf{x}_i)'\phi(\mathbf{x}_j)$, and $\hat{\boldsymbol{\alpha}}\in\mathbb{R}^N$ is the shrinked $\boldsymbol{\alpha}$ with its element $\hat{\alpha}_i=\frac{1}{|\mathcal{B}_l|}\sum_{c,m}\alpha_{l,c,m}$ if $i\in\mathcal{I}_l$. We can solve (27) approximately by enumerating the binary indicator vector $\mathbf{h}$ in a one bag by one bag fashion iteratively to maximize the objective value of (27) until $\mathbf{h}$ remains unchanged.

The whole algorithm for our weakly supervised domain generalization (WSDG) approach is listed in Algorithm 1.

[2]We initialize $\mathbf{h}_1$ by assigning 1 to its entries corresponding to the $\eta|\mathcal{B}_l|$ instances in each bag $\mathcal{B}_l$ with highest decision values, and 0 to other entries. Specifically, we first train the SVM classifiers by assigning the bag label to each instance and obtain the decision values of training instances by using the trained SVM classifiers.

Algorithm 1 Weakly Supervised Domain Generalization (WSDG) Algorithm

Input: Training data $\{(\mathcal{B}_l, Y_l)|_{l=1}^{L}\}$.

1: Initialize $t = 1$ and [2]$\mathcal{C} = \{\mathbf{h}_1\}$.
2: **repeat**
3: Set $t \leftarrow t + 1$.
4: Based on $\mathcal{H} = \mathcal{C}$, optimize the MKL problem in (18) to obtain $(\mathbf{d}, \boldsymbol{\alpha})$.
5: Find the violated $\mathbf{h}_t$ by solving (27) and set $\mathcal{C} \leftarrow \mathcal{C} \bigcup \mathbf{h}_t$.
6: **until** The objective of (18) converges.

Output: The learnt classifier $f(\mathbf{x})$.

4. Experiments

In this section, we demonstrate the effectiveness of our weakly supervised domain generalization (WSDG) approach for visual event recognition and image classification on three benchmark datasets. Moreover, we compare our WSDG-PI method with WSDG in order to demonstrate it is beneficial to utilize additional textual features as privileged information.

Experimental Settings: We evaluate our WSDG method for video event recognition and image classification by respectively using the crawled web videos and web images as training data.

For video event recognition, we use two benchmark datasets Kodak [35] and CCV [26] as the test sets. The Kodak dataset consists of 195 consumer videos from 6 event classes. The CCV dataset [26] consists of a training set of 4659 videos and a test set of 4658 videos from 20 semantic categories. Following [14], we only use the videos from the event related categories and merge the categories with similar semantic meanings. Finally, for the CCV dataset, we have 2440 videos from our classes as test samples.

To collect the training set, we crawl the web videos from the *Flickr* website by using those 6 (*resp.*, 5) event class names as queries for the Kodak (*resp.*, CCV) test set. We use the top 100 relevant web videos for each query to construct 20 bags for each class by uniformly partitioning those relevant videos based on their ranks, in which each bag contains 5 instances.

We extract the improved dense trajectory features for each video in Kodak/CCV/Flickr. Specifically, following [8], we use three types of space-time (ST) features (i.e., 96-dim Histogram of Oriented Gradient, 108-dim Histogram of Optical Flow and 192-dim Motion Boundary Histogram). We use the BoW representation for each type of ST features, in which the codebook is constructed by using k-means to cluster the ST features from all videos in the Flickr dataset into 2000 clusters. Finally, we concatenate the bag-of-words features from three types of descriptors to obtain a 6000-dim feature vector for each video.

For image classification, we use the BING dataset provided in [4] as the source domain and the benchmark Caltech-256 dataset as the test set. Following the setting in [24], we use the images from the first 30 categories in BING and Caltech-256. As suggested in [24], we use 20 training samples and 25 test samples for each category, so there are totally 600 training images and 750 test images. Similarly as for the Flickr dataset, we construct the bags for each category by uniformly partitioning the training images, in which each bag contains 5 instances. We use the DeCAF features [12], which has shown promising performance for image classification. Following [12], we use the outputs from the 6th layer as the visual features, which leads to $4,096$-dim DeCAF_6 features.

As the number of latent domains is unknown for web data, we set the number of latent domains as 2 for all methods as suggested in [24]. For our WSDG method, we empirically fix $C_1 = C_2 = 1$, $\eta = 0.2$ (*resp.*, 0.8) for video event recognition (*resp.*, image classification). For the baseline methods, we choose the optimal parameters according to their best accuracies on the test dataset.

Baselines: We compare our WSDG method with three groups of baseline methods, the domain generalization methods, the latent domain discovering methods, and the multi-instance learning (MIL) methods. The domain generalization methods include the domain-invariant component analysis (DICA) method [36] and the low-rank exemplar SVM (LRESVM) method [43]. The work in [27] can not be applied to our tasks, because we do not have the ground-truth domain labels for the training web datasets. The latent domain discovering methods include two methods in [24, 20] by using two strategies called "Match" and "Ensemble" as suggested in [43]. Specifically, the ensemble strategy is to learn the domain probabilities with the method in [24] and re-weight the decision values from different SVM classifiers corresponding to different latent domains. In the match strategy, we first use the MMD criterion to select the latent source domain, which is the most relevant one to the target domain, and then apply the SVM classifier corresponding to the most relevant latent source domain on the test samples. The MIL methods include two bag-level methods sMIL [6], KI-SVM [34] and two instance-level methods mi-SVM [2], MIL-CPB [31].

As the sub-categorization related methods are also relevant to our work, we further compare our work with the discriminative sub-categorization(Sub-Cate) method [23] and the MMDL method in [42].

In order to show the effectiveness for exploiting the latent domains and validate the MMD-based regularizer in (9), we also report the results of two simplified versions (referred to as WSDG_sim1 and WSDG_sim2) of our WSDG approach. In WSDG_sim1, we remove the MMD-based regularizer $\rho(\mathbf{B}, \mathbf{K} \circ (\mathbf{h}\mathbf{h}'))$ from our WSDG method by setting $C_2 = 0$. Based on WSDG_sim1, we further do

not consider the latent domain issues by setting the number of latent domains M=1 and call this special case as WSDG_sim2. In this case, we only have one source domain and our objective function in (9) reduces to that in (6).

Experimental Results: The experimental results are summarized in Table 1. We observe the domain generalization methods DICA and LRESVM, the sub-categorization methods Sub-Cate and MMDL, and the latent domain discovering methods [20] and [24] are generally better than SVM. The results demonstrate that it is beneficial to exploit the additional information like low-rank structure, subcategories, or latent domains in training web data for visual recognition by learning from web data.

We also observe the MIL baselines, including sMIL, mi-SVM, MIL-CPB, and KI-SVM, achieve better results than SVM on all three datasets, while those methods use different MIL assumptions for solving the MIL problem. Moreover, MMDL is better than the Sub-Cate method and MIL baselines, because it simultaneously uses the MIL technique for handling label noise in web data and exploits subcategories to learn more robust integrated classifiers.

Our special case WSDG_sim1 is better than the MIL baselines. One possible explanation is that we jointly learn the classifiers for multiple classes. WSDG_sim2 outperforms WSDG_sim1 on all three datasets, which shows it is useful to integrate multiple classifiers learnt from different latent domains. Our WSDG method further outperforms WSDG_sim2 on all three datasets, which validates the effectiveness of the MMD-based regularizer in (9). We also observe that WSDG_sim2 and WSDG outperform all the MIL methods [6, 34, 2, 31], which indicates that it is beneficial to use the discovered latent domains in our WSDG method to improve generalization capability of learnt classifiers to any unseen test domain. Moreover, WSDG_sim2 and WSDG are also better than the domain generalization methods DICA and LRESVM, which demonstrates their effectiveness for learning robust classifiers by handling label noise in the web images/videos .

Finally, our WSDG method achieves the best results on all three datasets, which clearly demonstrates the effectiveness of our method for video event recognition and image classification by learning from web data.

Utilizing Privileged Information: We use the Flickr web video dataset as the training set and the CCV and Kodak datasets as the test sets to evaluate our proposed WSDG-PI method discussed in Section 3.2.3. Note we cannot evaluate our WSDG-PI on the Caltech-256 dataset because textual information is not available for the Bing dataset provided in [4]. The tags associated with each Flickr video are crawled, from which we extract a $2,000$-dim term frequency (TF) feature for each video by using the $2,000$ most frequent words as the vocabulary. We also remove the stopwords before extracting the textual features. We use the textual features as privileged information because they are not available in the testing videos. The other experimental settings remain the same.

Table 1. Accuracies (%) of different methods for video event recognition and image classification. The best results are denoted in boldface.

Method	Testing Dataset		
	Kodak	CCV	Caltech-256
SVM [10]	34.36	40.84	70.93
sMIL [6]	38.46	41.34	71.33
mi-SVM [2]	37.95	46.38	71.47
MIL-CPB [31]	38.97	46.29	71.6
KI-SVM [34]	40.00	42.85	71.20
DICA [36]	42.05	44.10	70.80
LRESVM [43]	41.94	48.12	72.93
[24] (Match)	37.13	41.37	71.07
[24] (Ensemble)	37.42	41.40	70.08
[20] (Match)	40.93	44.44	71.47
[20](Ensemble)	42.39	47.51	72.40
Sub-Cate [23]	38.59	47.93	72.27
MMDL [42]	40.51	48.87	72.80
WSDG_sim1	42.56	47.47	71.87
WSDG_sim2	43.59	49.93	74.00
WSDG	**45.64**	**51.18**	**75.20**

Table 2. Accuracies (%) of different methods for video event recognition. The best results are denoted in boldface.

Method	Testing Dataset	
	Kodak	CCV
SVM+ [40]	36.92	45.05
sMIL-PI [33]	42.50	46.20
WSDG-PI (Ours)	**47.69**	**52.80**

In Table 2, we compare our WSDG-PI method with sMIL-PI, which outperforms other existing methods using privileged information as reported in [33]. We also include SVM+ [40] as a baseline method. After using the additional textual features, SVM+, sMIL-PI, and WSDG-PI achieve better results than the corresponding methods in Table 1 without using the textual features (*i.e.*, SVM, sMIL, and WSDG). Moreover, our WSDG-PI outperforms SVM+ and sMIL-PI, which again demonstrates it is beneficial to cope with label noise and simultaneously learn robust classifiers for better generalization capability.

5. Conclusion

In this paper, we have proposed a weakly supervised domain generalization (WSDG) approach for the visual recognition task by using loosely labeled web images/videos as training data. Our WSDG approach can handle label noise of training web images/videos and also enhance generalization capability of learnt classifiers to any unseen target domain. The effectiveness of our new approach has been demonstrated by the extensive experiments.

Acknowledgement

This work is supported by the Singapore MoE Tier 2 Grant (ARC42/13). This work is also supported by the Singapore National Research Foundation under its IDM Futures Funding Initiative and administered by the Interactive & Digital Media Programme Office, Media Development Authority.

References

[1] P.-A. Absil and A. L. Tits. Newton-KKT interior-point methods for indefinite quadratic programming. *Computational Optimization and Applications*, 36(1):5–41, 2007.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Proceedsing of the 25th Annual Conference on Neural Information Processing Systems*, pages 561–568, Vancouver, Canada, Dec. 2002.

[3] M. Baktashmotlagh, M. Harandi, and M. S. Brian Lovell. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the 14th International Conference on Computer Vision*, pages 769–776, Sydney, Australia, Dec. 2013.

[4] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Proceedsing of the 24th Annual Conference on Neural Information Processing Systems*, pages 181–189, Vancouver, Canada, Dec. 2010.

[5] L. Bruzzone and M. Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:770–787, May 2010.

[6] R. C. Bunescu and R. J. Mooney. Multiple instance learning for sparse positive bags. In *Proceedsing of the 24th IEEE International Conference on Machine Learning*, pages 105–112, Corvallis, OR, USA, Jun. 2007.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[8] L. Chen, L. Duan, and D. Xu. Event recognition in videos by learning from heterogeneous web sources. In *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2673, Portland, OR, Jun. 2013.

[9] L. Chen, W. Li, and D. Xu. Recognizing RGB images by learning from RGB-D data. In *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1418–1425, Columbus, OH, USA, Jun. 2014.

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[11] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[12] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proceedsing of the 31th IEEE International Conference on Machine Learning*, pages 647–655, Beijing, China, Jun. 2014.

[13] L. Duan, I. W. Tsang, and D. Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:465–479, Mar. 2012.

[14] L. Duan, D. Xu, and S.-F. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1338–1345, Providence, RI, 2012.

[15] L. Duan, D. Xu, and I. W. Tsang. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Trans. Neural Networks and Learning Systems*, 23(3):504–518, Mar. 2012.

[16] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1667–1680, Sep. 2012.

[17] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using second order information for training support vector machines. *The Journal of Machine Learning Research*, 6:1889–1918, 2005.

[18] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the 14th International Conference on Computer Vision*, pages 2960–2967, Sydney, Australia, Dec. 2013.

[19] P. V. Gehler and S. Nowozin. Infinite kernel learning. Technical report, Max Planck Institute for Biological Cybernetics, 2008.

[20] B. Gong, K. Grauman, and F. Sha. Reshaping visual datasets for domain adaptation. In *Proceedsing of the 27th Annual Conference on Neural Information Processing Systems*, pages 1286–1294, Lake Tahoe, Nevada, United States, Dec. 2013.

[21] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, Providence, RI, Jun. 2012.

[22] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proceedings of the 13th International Conference on Computer Vision*, pages 999–1006, Barcelona, Spain, Nov. 2011.

[23] M. Hoai and A. Zisserman. Discriminative subcategorization. In *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1666–1673, Portland, OR, Portland, OR 2013.

[24] J. Hoffman, K. Saeko, B. Kulis, and T. Darrell. Discovering latent domains for multisource domain adaptation. In *Proceedings of the 12th European Conference on Computer Vision*, pages 702–715. Firenze, Italy, Oct. 2012.

[25] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Scholkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608, Cancouver and Whistler, Canada, Dec. 2006.

[26] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 29, Trento, Italy, Apr. 2011.

[27] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *Proceedings of the 12th European Conference on Computer Vision*, pages 158–171. Firenze, Italy, Oct. 2012.

[28] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011.

[29] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1785–1792, Colorado Springs, CO, Jun. 2011.

[30] T. Leung, Y. Song, and J. Zhang. Handling label noise in video classification via multiple instance learning. In *Proceedings of the 13th International Conference on Computer Vision*, pages 2056–2063, Barcelona, Spain, Nov. 2011.

[31] W. Li, L. Duan, D. Xu, and I. W. Tsang. Text-based image retrieval using progressive multi-instance learning. In *Proceedings of the 13th International Conference on Computer Vision*, pages 2049–2055, Barcelona, Spain, Nov. 2011.

[32] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1134–1148, 2014.

[33] W. Li, L. Niu, and D. Xu. Exploiting privileged information from web data for image categorization. In *Proceedings of the 13th European Conference on Computer Vision*, pages 437–452. Zurich, Switzerland, Sep. 2014.

[34] Y.-F. Li, J. T. Kwok, I. W. Tsang, and Z.-H. Zhou. A convex method for locating regions of interest with multi-instance learning. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 15–30. Bled, Slovenia, Sep. 2009.

[35] A. C. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa. Kodak's consumer video benchmark data set. In *Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 245–254, University of Augsburg, Germany, Sep. 2007.

[36] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *Proceedsing of the 30th IEEE International Conference on Machine Learning*, pages 10–18, Atlanta, USA, Jun. 2013.

[37] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, Jul. 2014.

[38] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proceedings of the 11th European Conference on Computer Vision*, pages 213–226. Crete, Greece, Jun. 2010.

[39] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, Colorado Springs, CO, Jun. 2011.

[40] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

[41] S. Vijayanarasimhan and K. Grauman. Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In *Proceedings of the 21th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, AK, Jun. 2008.

[42] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu. Max-margin multiple-instance dictionary learning. In *Proceedsing of the 30th IEEE International Conference on Machine Learning*, pages 846–854, Atlanta, USA, Jun. 2013.

[43] Z. Xu, W. Li, L. Niu, and D. Xu. Exploiting low-rank structure from latent domains for domain generalization. In *Proceedings of the 13th European Conference on Computer Vision*, pages 628–643. Zurich, Switzerland, Sep. 2014.

[44] J. Yang, R. Yan, and A. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *Proceedings of the 15th international conference on Multimedia*, pages 188–197, Augsburg, Germany, Sep. 2007.

[45] D. Zhang, F. Wang, L. Si, and T. Li. M3IC: Maximum margin multiple instance clustering. In *Proceedsing of the 21st International Joint Conference on Artificial Intelligence*, volume 9, pages 1339–1344, Pasadena, CA, USA, Jul. 2009.

[46] M.-L. Zhang and Z.-H. Zhou. M3MIML: A maximum margin method for multi-instance multi-label learning. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 688–697, Pisa, Italy, Dec. 2008.