Exploiting Web Images for Dataset Construction: A Domain Robust Approach

Yazhou Yao, *Student Member, IEEE,* Jian Zhang, *Senior Member, IEEE,* Fumin Shen, *Member, IEEE,* Xiansheng Hua, *Fellow, IEEE,* and Zhenmin Tang

***Abstract*—Labelled image datasets have played a critical role in high-level image understanding. However, the process of manual labelling is both time consuming and labor intensive. To reduce the cost of manual labelling, there have been increasing research interests in automatically constructing image datasets by exploiting web images. However, datasets constructed by existing methods tend to have a weak domain adaptation ability, known as "dataset bias problem". To address this issue, in this work, we present a novel image dataset construction framework which can generalize well to unseen target domains. In specific, the given queries are first expanded by searching in the Google Books Ngrams Corpora to obtain a richer semantic description, from which the visually non-salient and less relevant expansions are filtered out. By treating each unfiltered expansion as a "bag" and the retrieved images therein as "instances", we then formulate images selection as a Multi-instance learning problem with constrained positive bags. By this approach, images from different distributions will be kept while with noisy images filtered out. To verify the effectiveness of our proposed approach, we build a domain robust image dataset with 20 categories (referred as DRID-20). We compare the image classification ability, cross-dataset generalization ability and dataset diversity of DRID-20 with three publicly available datasets STL-10, CIFAR-10 and ImageNet. The experimental results indicated the domain robustness of our dataset. In order to further compare with other weak and web supervised baseline methods, we run object detection on PASCAL VOC 2007 using our data. The results demonstrated our method is superior to the weak and web supervised state-of-the-art methods on object detection.**

***Index Terms*—Domain robust, multiple query expansions, image dataset construction, MIL**

I. Introduction

The availability of labelled image datasets has proven invaluable for high-level image understanding. For example, ImageNet [1] has acted as one of the most important factors in the recent advance of developing and deploying visual representation learning models (e.g., deep CNN). However, the process of constructing ImageNet is both time consuming and labor intensive. Recently, there have been increasing research interests in automatically constructing image datasets by exploiting web images[8], [19], [24], [35]. Existing methods [8], [19], [35] usually take an iterative mechanism in the process of images selection. Due to the visual feature distributions of images selected by the iterative mechanism, these datasets tend to have statistical problems, which are known as the dataset bias problem [17], [21], [38]. Fig. 1 shows the "airplane" images from four different image datasets. We observe some significant differences: PASCAL [6] have "airplanes" at flying view-points, while SUN [30] tend to have distant views in the airport; Caltech [28] has a strong preference for side views and ImageNet [1] is rich in diversity, but mainly in close-range views. Classifiers learned from these datasets usually have poor performance on domain adaptation [17]. To address this problem, a large number of domain adaptation approaches have been proposed for various vision tasks [22], [27], which explicitly coped with noisy labels of web images. Specifically, the images are partitioned into a set of clusters and each cluster is treated as a "bag" with the images in each bag as "instances". As a result, these tasks can be formulated as a Multi-instance learning (MIL) problem and different MIL methods were proposed in [22], [27]. However, for all of these methods, the yield is limited by the restriction of diversity which provided by the image search engine with one query.

Fig. 1: Most discriminative images from 4 different datasets.

In order to obtain high accuracy and diversity candidate images, as well as to overcome the download restriction of image search engine, [5], [24] propose to use multiple query expansions instead of one query in the process of collecting candidate images from image search engine. The problem is

Y. Yao and J. Zhang are with the Global Big Data Technologies Center, University of Technology Sydney, NSW 2007, Australia.

F. Shen is a Lecturer in School of Computer Science and Engineering, University of Electronic Science and Technology of China.

X. Hua is a researcher/senior director in Alibaba Group, Hangzhou, China.

Z. Tang is a Professor in School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.

Manuscript received ; revised .

these methods still take iterative mechanisms in the process of images selection which leads to a dataset bias problem [17], [21], [38].

Based on these motivations, in this work, we are targeting at constructing image dataset in a scalable way while ensuring robustness and accuracy. The basic idea is to leverage multiple query expansions for initial candidate images collection and MIL based methods for selecting images from different distributions. In order to obtain multiple query expansions, we expand query to a set of query expansions and then most of the noisy expansions are filtered out. After we obtain the raw image dataset with unfiltered query expansions, MIL based methods are applied to filter individual and group noisy images. To verify the effectiveness of our proposed approach, we build an image dataset with 20 categories. We compare the image classification ability, cross-dataset generalization ability and dataset diversity of our DRID-20 with three manually labelled image datasets CIFAR-10, STL-10 and ImageNet to demonstrate the domain robustness of our dataset. Besides, we also report the results of object detection on PASCAL VOC 2007, we then compare the object detection ability of our method with four baseline methods. Our contributions are threefold:

[1.] To the best of our knowledge, we are the first proposal of constructing domain robust image dataset automatically. Our proposed approach based on multiple query expansions and Multiple-instance learning considers the source of candidate images and keep images from different distributions. Hence, the dataset constructed by our approach can efficiently ease the dataset bias problem.

[2.] The proposed framework is a generalized one which makes constructing domain robust image dataset while ensuring accuracy feasible. Several experiments have demonstrated that our dataset DRID-20 has better image classification ability, cross-dataset generalization ability and diversity. Additionally, it is worth mentioning that DRID-20 shows domain robustness without manual labelling.

[3.] We have released image dataset DRID-20. We hope the diversity of DRID-20 can offer unparalleled opportunities to researchers in the Multi-instance learning, transfer learning, image dataset construction and other related fields.

This paper is an extended version of [38]. The extensions include: Considering both of bag level and instance level noisy images instead of just instance level noisy images in the process of images selection, we took combination of bag level and instance level selection mechanisms and achieved a better results; Comparing the image classification ability and dataset diversity of our dataset DRID-20 with STL-10, CIFAR-10 and ImageNet; Increasing the categories in dataset from 10 to 20, and then our dataset DRID-20 cover all categories in PASCAL VOC dataset.

The rest of the paper is organized as follows: In Section 2, a brief discussion of related works are given. The proposed algorithm including query expanding, noisy expansions filtering and noisy images filtering is described in Section 3. We evaluate the performance of the proposed algorithm with several other methods in Section 4. Finally the conclusion and future work are offered in Section 5.

II. Related works

Considering the importance of labelled image datasets in the area of high-level image understanding, lots of efforts have been involved in image dataset construction. In general, these efforts can be divided into three principal categories: manual based methods, semi-automatic based methods and automatically based methods.

A. Manual and Semi-automatic Based Methods

In the early years, manual labelling is the most important way to construct image datasets. (e.g., STL-10 [25], CIFAR-10 [11], PASCAL VOC [6], ImageNet [1] and Caltech-101 [28]). The process of constructing these datasets is mainly by submitting keywords to image search engine to download candidate images, then cleaning these candidate images by manual annotation. This method has a high accuracy, but is labor intensive.

In order to reduce the cost of manual labelling, some works also focused on active learning (a special case of semi-supervised method)[29][32][33]. [29] randomly label some seed images to learn visual classifiers. Then the learned visual classifiers are implemented to do image classifications on unlabelled images, finding out un-confident images for manual labelling. Here un-confident images are those whose probability are classified into positive and negative close to 0.5. The process is iterated until sufficient classification accuracy is achieved. [32] presented an active learning framework to simultaneously learn contextual models for scene understanding tasks (multi-class classification). [33] presented an approach for on-line learning of object detectors, in which the system automatically refines its models by actively requesting crowd-sourced annotations on images crawled from the web. However, both of manual labelling and active learning require pre-existing annotations which often results in one of the biggest limitations to develop a large scale image dataset.

B. Automatically Based Methods

To further decrease the cost of manual annotation, automatic methods have attracted more and more people's attention [19], [8], [24], [35]. [19] adopt text information to re-rank images retrieved from web search and used these top-ranked images to learn visual models to re-rank images once again. The advantage of these methods is eliminating the need for manual intervention. [35] leveraged the first few images returned from image search engine to train image classifier, classifying images as positive or negative. When the image is classified as a positive sample, the classifier uses incremental learning to refine its model. With the increase of classifier accepting more positive images, the trained classifier will reach a robust level for this query. [8] proposed to use clustering based method to filter "group" noisy images and propagation based method to filter individual noisy images. However, for methods [19], [8], [35], the domain adaptation ability is limited by the restriction of initial candidate images and iterative mechanism in the process of image selection. In order to obtain high diversity candidate images, [24] proposed to use multiple query expansions instead of one query in the process of initial

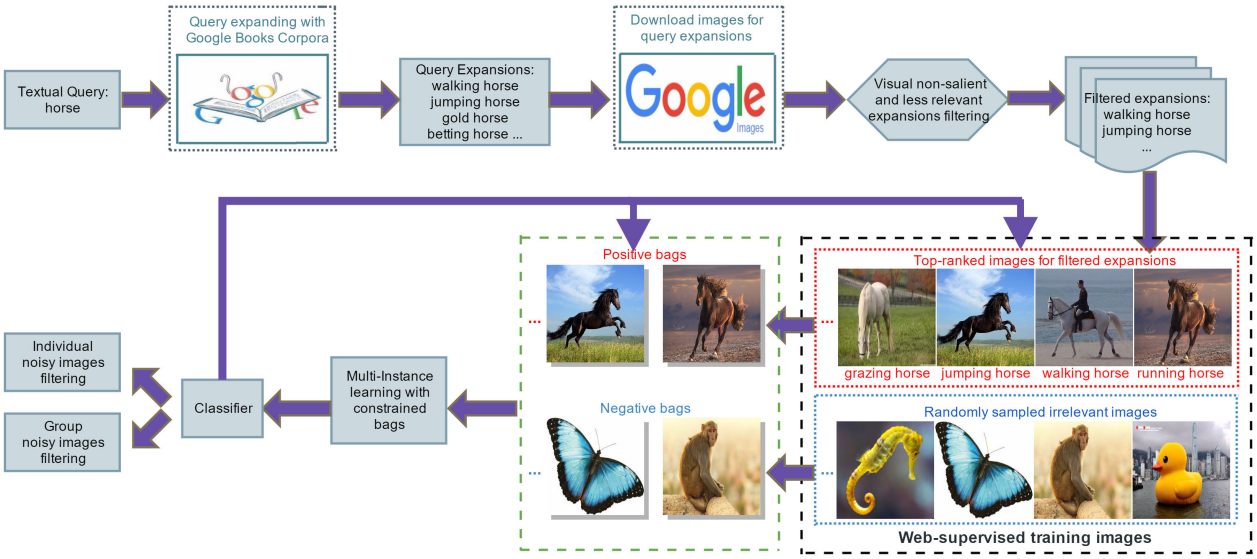

Fig. 2: Flowchart of the proposed approach.

candidate images collection. Then take iterative mechanism for noisy images filtering. However, these previous automatic works mainly focus on accuracy and scale in the process of image dataset construction, which often results in a poor performance on domain adaptation.

C. Other Related Works

There are lots of work related to the step of query expansions generating and noisy images filtering though not aiming at image dataset construction. Since most image search engine restricts the image numbers returned for each query, WordNet [16] and ConceptNet [36] are often used to obtain synonyms to overcome the download restriction of image search engine. The advantage of WordNet and ConceptNet is synonyms are usually relevant to the given query. We almost do not need to purify these synonyms. However, WordNet and ConceptNet are usually not rich enough for query expanding. What's worse, the images returned from image search engine using synonyms tend to have the homogenization problem which will result in poor performance on domain adaptation. Recent works [24], [5] proposed to use Google Books Ngrams Corpora (GBNC) [15] to expand query to a set of query expansions. Google Books Ngrams Corpora cover almost all related queries at the text level. It's much general and richer than WordNet and ConceptNet. The disadvantage of using GBNC for query expanding is it may also brings some noisy query expansions. Recently, word embedding [4], [37] provides a learning based method to compute the word-word similarity distance which can be used to filter noisy query expansions. In this paper, we use GBNC to expand query to a set of query expansions, and then take both word-word and visual-visual similarity distance to filter noisy query expansions.

In order to efficiently ease the dataset bias problem, some authors have developed domain adaptation approaches for vision tasks. [27] clustered relevant images using both textual and visual features. By treating each cluster as a "bag" and the images in the bag as "instances", the authors formulated this problem as a Multi-instance learning problem (MIL) which learns a target decision function for images re-ranking. However, the yield is limited by the restriction of initial candidate images which were obtained using one query from the Internet. In this paper, we focus on Multi-instance learning (MIL) based method, as it can keep images from different data distribution while noisy images filtered out.

In summary, existing automatically based methods reduce the cost of manual annotation by leveraging the generalization ability of machine learning models. However, the generalization ability is affected by both the quality of initial candidate images and the capability of models to keep images from different distributions. In other words, previous works mainly focus on accuracy and scale. Most of them take an iterative mechanism in the process of images selection which often results in a dataset bias problem. To the best of our knowledge, we are the first proposal of constructing domain robust image dataset automatically. We gain the domain adaptation ability of our dataset by maximizing both of the initial candidate images and the final selected images from different data distributions.

III. Proposed Approach

We seek to construct a domain robust image dataset which can generalize well to unseen target domains. As shown in Fig. 2, we propose our web supervised image dataset construction framework by three major steps: query expanding, noisy expansions filtering and noisy images filtering. Specifically, by searching in the GBNC [15], a set of semantically rich expansions are obtained, from which the visually non-salient and less relevant expansions are then filtered by exploiting both word-word and visual-visual similarity. After we obtain the candidate images by retrieving these unfiltered expansions with image search engine, we treat each expansion as a "bag"

and the images in each bag as "instances". We then formulate this task as a MIL problem with constrained positive bags. By this approach, images from different data distributions will be kept while noisy images filtered out, and a domain robust image dataset will be constructed.

A. Query Expanding

Image datasets constructed by existing methods tend to have a higher accuracy, but usually have weak domain adaptation ability [17], [21], [38]. In order to construct a domain robust image dataset, we expand query (e.g., "horse") to a set of query expansions (e.g., "jumping horse, walking horse, roaring horse", etc.) and then use these different query expansions (corresponding images) to reflect different "visual patterns" of the query. We use GBNC to discover query expansions for the given query with Parts-Of-Speech (POS), specifically with NOUN, VERB, ADJECTIVE and ADVERB. Our motivation is to identify all related query expansions. GBNC is much more general and richer than WordNet [16] and ConceptNet [36]. Using GBNC can help us find all expansions for any possible query the human race has ever written down in books.

B. Noisy Query Expansions Filtering

Through query expanding, we obtain a rich semantic description for the given query. However, query expanding not only brings all useful query expansions, but also some noisy query expansions. These noisy query expansions can be roughly divided into two types: (1) visual non-salient (e.g., "betting horse") and (2) less relevant (e.g., "sea horse"). Using these noisy query expansions to retrieve images will result in a bad influence on dataset accuracy and robustness.

1) Visual non-salient expansions filtering: From the perspective of visual, we want to identify visually salient query expansions and eliminate visual non-salient query expansions in this step. The intuition is that visually salient expansions should exhibit predictable visual patterns. Hence, we use image-classifier based filtering method. For each query expansion, we directly download the top N images from Google image search engine as positive images (based on the fact that the top few images returned from image search engine tend to be positive); then randomly split these images into a training set and validation set $\mathrm{I_i} = \{\mathrm{I_i^t}, \mathrm{I_i^v}\}$, we gather a random pool of negative images and split them into a training set and validation set $\bar{\mathrm{I}} = \{\bar{\mathrm{I}}^{\mathrm{t}}, \bar{\mathrm{I}}^{\mathrm{v}}\}$. We train a linear SVM classifier $\mathrm{C_i}$ with $\mathrm{I_i^t}$ and $\bar{\mathrm{I}}^{\mathrm{t}}$ using dense HOG features. Then we use $\{\mathrm{I_i^v}, \bar{\mathrm{I}}^{\mathrm{v}}\}$ as validation images to calculate the classification results. We declare a query expansion i to be visually salient if the classification results $\mathrm{S_i}$ giving a relatively high score.

2) Less relevant expansions filtering: From the perspective of relevance, we want to identify both semantic and visual relevant expansions for the given query. The intuition is that relevant expansions should exhibit a relatively small semantic and visual distance. Therefore, we use combined word-word and visual-visual similarity distance based filtering method.

Words and phrases acquire meaning from the way they are employed in society. For computers, the equivalent of "society" is "database", and the equivalent of "use" is "a way to search the database" [4]. Normalized Google Distance (NGD) constructs a method to extract semantic similarity distance from the World Wide Web (WWW) using Google page counts [4]. For a search term x and search term y, NGD is defined by:

$$\mathrm{NGD(x,y)} = \frac{\max\{\mathrm{logf(x)}, \mathrm{logf(y)}\} - \mathrm{logf(x,y)}}{\mathrm{logN} - \min\{\mathrm{logf(x)}, \mathrm{logf(y)}\}} \quad (1)$$

where $\mathrm{f(x)}$ denotes the number of pages containing x, $\mathrm{f(x,y)}$ denotes the number of pages containing both x and y and N is the total number of web pages searched by Google.

We denote the semantic distance of all query expansions by a graph $\mathrm{G_g} = \{\mathrm{N}, \mathrm{D}\}$ where each node represents a query expansion and its edge represents the NGD between the two nodes. We set the target query as center $\mathrm{(x)}$ and other query expansions have a score $(\mathrm{D_{xy}})$ which corresponds to the distance to the target query. Similarly, we represent the visual distance of query and expansions by a graph $\mathrm{G_v} = \{\mathrm{C}, \mathrm{E}\}$ where each node represents a query expansion and each edge represents the visual distance between query and expansions. The feature is 1000 dimensional bag of visual words based on SIFT features. The edge weight $\mathrm{E_{xy}}$ corresponds to the Euclidean distance.

The semantic distance and visual distance will be used to construct a new 2 dimensional feature $\mathrm{V} = [\mathrm{D_{xy}}; \mathrm{E_{xy}}]$. Then the problem is to calculate the importance weight w and bias penalty b in decision function $\mathrm{f(x)} = \mathrm{w^T x} + \mathrm{b}$ to determine whether the expansion is relevant or not. There are lots of methods to obtain these coefficients w and b. Here we take linear SVM to work around this problem. Although linear SVM is not the prevailing state-of-the-art method for classification, we found our method to be effective in pruning irrelevant query expansions.

Unfiltered expansions are then used to retrieve the top M images from image search engine to construct the raw image dataset. Regardless of the fact that our method is not able to remove noisy expansions thoroughly in most of the cases, the raw image dataset constructed by our method still achieves a much higher accuracy than directly using the Flickr or Google image data. Besides, the raw image dataset constructed through unfiltered query expansions has a much richer visual patterns.

C. Noisy Images Filtering

Although Google image search engine has ranked the returned images, some noisy images are still included. In addition, a few unfiltered noisy expansions will also bring noisy images to the raw image dataset. In general, these noisy images can be divided into two types: group noisy images (caused by unfiltered noisy expansions) and individual noisy images (due to the error index of image search engine). In order to filter these group and individual noisy images, as well as to keep images from different distributions, we take Multi-instance learning (MIL) based methods instead of iterative mechanism in the process of noisy images filtering.

By treating each unfiltered expansion as a "bag" and the images corresponding to the expansion as "instances", we formulate a Multi-instance learning problem by selecting a

subset of bags and a subset of images from each bag to construct the domain robust image dataset for the given query. Since the precision of images returned from Google image search engine tends to have a relatively high accuracy, we define each positive bag at least have a portion of δ positive instances which can effectively filter group noisy images caused by unfiltered noisy query expansions.

We denote each instance as $\mathrm{x_i}$ with its label $\mathrm{y_i} \in \{0,1\}$, where i=1,...,n. We also denote the label of each bag $\mathrm{B_I}$ as $\mathrm{Y_I} \in \{0,1\}$. The transpose of a vector or matrix is represented by superscript $'$ and the element-wise product between two matrices is represented by $\odot$. Moreover, we define the identity matrix as $\mathbf{I}$ and $\mathbf{0}$, $\mathbf{1} \in \Re^n$ denote the column vectors of all zeros and ones, respectively. The inequality $\mathbf{u} = [\mathrm{u_1}, \mathrm{u_2}...\mathrm{u_n}]' \geq \mathbf{0}$ means that $\mathrm{u_i} \geq 0$ for i=1,...,n.

1) Individual noisy images filtering: For individual noisy images filtering, the decision function is assumed in the form of $\mathrm{f(x)} = \mathrm{w}'\varphi(\mathrm{x}) + \mathrm{b}$ and it is to be learned from the raw image dataset. We employ the formulation of Lagrangian SVM, in which the square bias penalty b^2 and the square hinge loss for each instance are used in the objective function. Then the decision function can be learned by minimizing the following structural risk function:

$$\min_{\mathbf{y},\mathrm{w},\mathrm{b},\rho,\varepsilon_\mathrm{i}} \quad \frac{1}{2}\left(\|\mathrm{w}\|^2 + \mathrm{b}^2 + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{n}} \varepsilon_\mathrm{i}^2\right) - \rho \tag{2}$$

$$\text{s.t.} \quad \mathrm{y_i}(\mathrm{w}'\varphi(\mathrm{x_i}) + \mathrm{b}) \geq \rho - \varepsilon_\mathrm{i}, \mathrm{i} = 1, ...\mathrm{n}. \tag{3}$$

$$\sum_{\mathrm{i}:\mathrm{x_i}\in \mathrm{B_I}} \frac{\mathrm{y_i}+1}{2} \geq \delta\,|\mathrm{B_I}| \quad \text{for} \quad \mathrm{Y_I} = 1, \qquad \mathrm{y_i} = 0 \quad \text{for} \quad \mathrm{Y_I} = 0 \tag{4}$$

where φ is a mapping function that maps x from the original space into a high dimensional space $\varphi(\mathrm{x})$, $\mathrm{C} > 0$ is a regularization parameter and ε_i values are slack variables. The margin separation is defined as $\rho / \|\mathrm{w}\|$. $\mathrm{y} = [\mathrm{y_1}...\mathrm{y_n}]'$ means the vector of instance labels, $\lambda = \{\mathbf{y} | \mathrm{y_i} \in \{0,1\}\}$ and $\mathbf{y}$ satisfies constraint (4). By introducing a dual variable α_i for inequality constraint (3) and kernel trick $\mathrm{k_{ij}} = \varphi(\mathrm{x_i})'\varphi(\mathrm{x_j})$, we arrive at the below optimization problem:

$$\min_{\mathbf{y}\in\lambda} \max_{\alpha} \quad -\frac{1}{2}\Big(\sum_{\mathrm{i,j}=1}^{\mathrm{n}} \alpha_\mathrm{i}\alpha_\mathrm{j}\mathrm{y_i}\mathrm{y_j}\mathrm{k_{ij}} + \sum_{\mathrm{i,j}=1}^{\mathrm{n}} \alpha_\mathrm{i}\alpha_\mathrm{j}\mathrm{y_i}\mathrm{y_j} + \frac{1}{\mathrm{C}}\Big) \tag{5}$$

where $\alpha_\mathrm{i} \geq 0$, $\sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_\mathrm{i} = 1$ and $\alpha = [\alpha_1, \alpha_2...\alpha_n]'$. By defining $\mathbf{K} = [\mathrm{k_{ij}}]$ as a $\mathrm{n}\times\mathrm{n}$ kernel matrix, $\nu = \{\alpha | \alpha \geq \mathbf{0}, \alpha'\mathbf{1} = 1\}$ and $\widetilde{\mathbf{K}} = \mathbf{K} + \mathbf{1}\mathbf{1}'$ as a $\mathrm{n}\times\mathrm{n}$ transformed kernel matrix for the augmented feature mapping $\widetilde{\varphi}(\mathrm{x}) = [\varphi(\mathrm{x})'. \quad 1]'$ of kernel $\widetilde{\mathrm{k_{ij}}} = \widetilde{\varphi}(\mathrm{x_i})'\widetilde{\varphi}(\mathrm{x_j})$. (5) can be rewritten as follows:

$$\min_{\mathbf{y}\in\lambda} \max_{\alpha\in\nu} \quad -\frac{1}{2}\alpha'(\widetilde{\mathbf{K}} \odot \mathbf{y}\mathbf{y}' + \frac{1}{\mathrm{C}}\mathbf{I})\alpha \tag{6}$$

(6) is a mixed integer programming problem with respect to the instance labels $\mathrm{y_i} \in \{0,1\}$. We take Label-Generating MMC (LG-MMC) algorithm which was proposed in [14] to solve this mixed integer programming problem. We firstly consider interchanging the order of $\max_{\alpha\in\upsilon}$ and $\min_{\mathbf{y}\in\lambda}$ in (6) and get:

$$\max_{\alpha\in\nu} \min_{\mathbf{y}\in\lambda} \quad -\frac{1}{2}\alpha'(\widetilde{\mathbf{K}} \odot \mathbf{y}\mathbf{y}' + \frac{1}{\mathrm{C}}\mathbf{I})\alpha. \tag{7}$$

According to the minmax theorem [10], the optimal objective of (6) is an upper bound of (7). We further rewrite (7) as:

$$\max_{\alpha\in\nu}\left\{\max_{\vartheta} -\theta | \theta \geq \frac{1}{2}\alpha'(\widetilde{\mathbf{K}} \odot \mathbf{y}^\mathrm{t}\mathbf{y}^{\mathrm{t}'} + \frac{1}{\mathrm{C}}\mathbf{I})\alpha, \forall \mathbf{y}^\mathrm{t} \in \lambda\right\} \tag{8}$$

$\mathbf{y}^t$ is any feasible solution in λ. For the inner optimization sub-problem, let $\mathrm{u_t} \geq 0$ be the dual variable for inequality constraint. Its Lagrangian can be obtained as:

$$-\theta + \sum_{\mathrm{t}:\mathbf{y_t}\in\lambda} \mathrm{u_t}(\theta - \frac{1}{2}\alpha'(\widetilde{\mathbf{K}} \odot \mathbf{y}^\mathrm{t}\mathbf{y}^{\mathrm{t}'} + \frac{1}{\mathrm{C}}\mathbf{I})\alpha). \tag{9}$$

Setting the derivative of (9) with respect to θ to zero, we have $\sum \mathrm{u_t} = 1$. $\mathbf{M} = \{\mathbf{u} | \sum \mathrm{u_t} = 1, \mathrm{u_t} \geq 0\}$ is denoted as the domain of $\mathbf{u}$, where $\mathbf{u}$ is the vector of $\mathrm{u_t}$. Then the inner optimization sub-problem is replaced with its dual and (8) can be rewritten as:

$$\max_{\alpha\in\nu} \min_{\mathbf{u}\in\mathbf{M}} \quad -\frac{1}{2}\alpha'(\sum_{\mathrm{t}:\mathbf{y}^\mathrm{t}\in\lambda} \mathrm{u_t}\widetilde{\mathbf{K}} \odot \mathbf{y}\mathbf{y}' + \frac{1}{\mathrm{C}}\mathbf{I})\alpha$$

or

$$\min_{\mathbf{u}\in\mathbf{M}} \max_{\alpha\in\nu} \quad -\frac{1}{2}\alpha'(\sum_{\mathrm{t}:\mathbf{y}^\mathrm{t}\in\lambda} \mathrm{u_t}\widetilde{\mathbf{K}} \odot \mathbf{y}\mathbf{y}' + \frac{1}{\mathrm{C}}\mathbf{I})\alpha. \tag{10}$$

Here, we can interchange the order of $\max_{\alpha\in\nu}$ and $\min_{\mathbf{u}\in\mathbf{M}}$ because the objective function is concave in α and convex in $\mathbf{u}$. Besides, (10) can be regarded as multiple kernel learning (MKL) problem [2], the target kernel matrix is a convex combination of base kernel matrices $\left\{\widetilde{\mathbf{K}} \odot \mathbf{y_t}\mathbf{y_t}'\right\}$. Although λ is finite and (10) is an MKL problem, we can't directly use existing MKL techniques like [18] to solve this problem. The reason is that the exponential number of possible labellings $\mathbf{y}_\mathrm{t} \in \lambda$ and the base kernels is also exponential in size makes direct MKL computes intractable.

Fortunately, not all the constraints in (8) are active at optimality, we can employ cutting-plane algorithm [9] to find a subset $\zeta \in \lambda$ of the constraints that can well approximate the original optimization problem. The detail solutions of cutting-plane algorithm for (10) are described in Algorithm 1. Finding most violated constraint $\mathbf{y}^t$ is the most challenging part in the cutting-plane algorithm. According to (5), the most violated $\mathbf{y}^t$ is equivalent to the following optimization problem:

$$\max_{\mathbf{y}\in\lambda} \quad \sum_{\mathrm{i,j}=1}^{\mathrm{n}} \alpha_\mathrm{i}\alpha_\mathrm{j}\mathrm{y_i}\mathrm{y_j}\mathrm{k_{ij}}. \tag{11}$$

Algorithm 1 Cutting-plane algorithm for (10)

1: Initialize $\mathrm{y_i} = \mathrm{Y_I}$ for $\mathrm{x_i} \in \mathrm{B_I}$ as $\mathbf{y}^1$, and set $\zeta = \{\mathbf{y}^1\}$;
2: Use MKL to solve α and $\mathbf{u}$ in (10) with ζ;
3: Select most violated $\mathbf{y}^t$ with α and set $\zeta = \mathbf{y}^t \cup \zeta$;
4: Repeat step 2 and step 3 until convergence.

We solve this integer optimization problem by enumerating all possible candidates of $\mathbf{y}^t$. Here we only enumerate the possible labelling candidates of the instances in positive bags as all instances in the negative bags are assumed to be negative in our paper. Finally, we can derive the decision function from the raw image dataset for the given query as:

$$\mathrm{f}(\mathrm{x}) = \sum_{\mathrm{i}:\alpha_\mathrm{i}\neq 0} \alpha_\mathrm{i}\widetilde{\mathrm{y}}_\mathrm{i}\widetilde{\mathrm{k}}(\mathrm{x},\mathrm{x}_\mathrm{i}) \tag{12}$$

where $\widetilde{\mathrm{y}}_\mathrm{i} = \sum_{\mathrm{t}:\mathbf{y}^\mathrm{t}\in\lambda} \mathrm{u}_\mathrm{t}\mathrm{y}_\mathrm{i}^\mathrm{t}$ and $\widetilde{\mathrm{k}}(\mathrm{x},\mathrm{x}_\mathrm{i}) = \mathrm{k}(\mathrm{x},\mathrm{x}_\mathrm{i}) + 1$. The decision function will be used to filter individual noisy images in each bag which corresponding to unfiltered query expansions.

2) Group noisy images filtering: For group noisy images filtering (caused by unfiltered noisy expansions), we represent bag $\mathrm{B_I}$ with compound feature $\phi_{\mathrm{f,k}}$ of its first k positive instances:

$$\phi_{\mathrm{f,k}}(\mathrm{B_I}) = \frac{1}{\mathrm{k}} \sum_{\mathrm{x_i}\in\Psi^*_{\mathrm{f,k}}(\mathrm{B_I})} \mathrm{x_i} \tag{13}$$

with

$$\Psi^*_{\mathrm{f,k}}(\mathrm{B_I}) = \underset{\Psi\subseteq\mathrm{B_I},|\Psi|=\mathrm{k}}{\operatorname{argmax}} \sum_{\mathrm{x_i}\in\Psi} \mathrm{f}(\mathrm{x_i}). \tag{14}$$

We refer to the instances in $\Psi^*_{\mathrm{f,k}}(\mathrm{B_I})$ as the first k instances of $\mathrm{B_I}$ according to classifier f (see Equation 12). Since the closer of images in $\mathrm{B_I}$ from the bag centre, the higher probability of these images to be relevant to the bag. The assignment of relatively heavier weights to images which have short distance to bag centre would increase the accuracy of classifying bag $\mathrm{B_I}$ to be positive or negative, then increase the efficiency of filtering noisy group images. Following [39], we assume $\xi_\mathrm{i} = [1+\exp(\alpha \log \mathrm{d}(\mathrm{x_i}) + \beta)]^{-1}$ be a weighting function, $\mathrm{d}(\mathrm{x_i})$ represents the Euclidean distance of images $\mathrm{x_i}$ to the bag centre, $\alpha \in \mathbb{R}_{++}$ and β are scaling and offset parameters which can be determined by cross-validation. The representation of (13) for bag $\mathrm{B_I}$ can be generalized to a weighted compound feature:

$$\phi_{\mathrm{f,k}}(\mathrm{B_I}) = \phi(\mathrm{X},\mathrm{h}^*) = \frac{\mathrm{Xh}^*}{\xi^\mathrm{T}\mathrm{h}^*} \tag{15}$$

with

$$\mathrm{h}^* = \underset{\mathrm{h}\in\mathrm{H}}{\operatorname{argmax}}\ \mathrm{f}\left(\frac{\mathrm{Xh}}{\xi^\mathrm{T}\mathrm{h}}\right), \quad \mathrm{s.t.} \quad \sum_\mathrm{i} \mathrm{h_i} = \mathrm{k} \tag{16}$$

where $\mathrm{X} = [\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3.., \mathrm{x_i}] \in \mathbb{R}^{\mathrm{D}\times\mathrm{i}}$ is a matrix whose columns are the instances of bag $\mathrm{B_I}$, $\xi = [\xi_1, \xi_2, \xi_3 ... \xi_\mathrm{i}]^\mathrm{T} \in \mathbb{R}^\mathrm{i}_{++}$ are the vectors of weights, and $\mathrm{h}^* \in \mathrm{H} = \{0,1\}^\mathrm{i} \setminus \{0\}$ ($\sum_\mathrm{i} \mathrm{h_i} = \mathrm{k}$) is an indicator function for the first k positive instances of bag $\mathrm{B_I}$.

Then classifying rule of bag $\mathrm{B_I}$ to be positive or negative is:

$$\mathrm{f}_\omega(\mathrm{X}) = \underset{\mathrm{h}\in\mathrm{H}}{\max}\ \omega^\mathrm{T}\phi(\mathrm{X},\mathrm{h}), \quad \sum_\mathrm{i} \mathrm{h_i} = \mathrm{k} \tag{17}$$

where $\omega \in \mathbb{R}^\mathrm{D}$ is the vector of classifying coefficients, $\phi(\mathrm{X},\mathrm{h}) \in \mathbb{R}^\mathrm{D}$ is the feature vector of (15), h is a vector of latent variables and H is the hypothesis space $\{0,1\}^\mathrm{i} \setminus \{0\}$. The learning problem is to determine the parameter vector ω.

Algorithm 2 Concave-convex procedure for solving (21)

1: Initialize ω with SVM by setting $\mathrm{h} = \mathbf{1} \in \mathbb{R}^\mathrm{i}$;
2: Compute a convex upper bound using the current model for the second term of (21);
3: Minimize this upper bound by solving a structural SVM problem via the proximal bundle method [34];
4: Repeat step 2 and step 3 until convergence.

Given a training set $\tau = \{\mathrm{B_I}, \mathrm{Y_I}\}_{\mathrm{I}=1}^\mathrm{n}$, this is a latent SVM learning problem:

$$\min_\omega \frac{1}{2}\|\omega\|^2 + \mathrm{C}\sum_{\mathrm{I}=1}^\mathrm{n} \max\left(0, 1 - \mathrm{Y_I}\mathrm{f}_\omega\left(\mathrm{X_{B_I}}\right)\right). \tag{18}$$

Before solve (18), we firstly solve the classifying rule of (17). We need to solve the below problem:

$$\underset{\mathrm{h}\in\mathrm{H}}{\max} \frac{\omega^\mathrm{T}\mathrm{Xh}}{\xi^\mathrm{T}\mathrm{h}}, \quad \mathbf{s.t.}, \sum_\mathrm{i} \mathrm{h_i} = \mathrm{k}. \tag{19}$$

This is an integer linear-fractional programming problem. Since $\xi \in \mathbb{R}^\mathrm{i}_{++}$, (19) is identical to the relaxed problem:

$$\underset{\mathrm{h}\in ß^\mathrm{i}}{\max} \frac{\omega^\mathrm{T}\mathrm{Xh}}{\xi^\mathrm{T}\mathrm{h}}, \quad \mathbf{s.t.}, \sum_\mathrm{i} \mathrm{h_i} = \mathrm{k} \tag{20}$$

where $ß^\mathrm{i} = [0,1]^\mathrm{i}$ is a unit box in $\mathbb{R}^\mathrm{i}$. (20) is a linear-fractional programming problem and can be reduce to a linear programming problem of $\mathrm{i}+1$ variables and $\mathrm{i}+2$ constraints [40].

In this work, we take concave-convex procedure (CCCP) [41] to solve (18). We rewrite the objective of (18) as two convex functions:

$$\begin{aligned}\min_\omega \Bigg[\frac{1}{2}\|\omega\|^2 + \mathrm{C}\sum_{\mathrm{I}\in\mathrm{D_n}} \max\left(0, 1 + \mathrm{f}_\omega\left(\mathrm{X_{B_I}}\right)\right) + \\ \mathrm{C}\sum_{\mathrm{I}\in\mathrm{D_p}} \max\left(\mathrm{f}_\omega\left(\mathrm{X_{B_I}}\right), 1\right)\Bigg] - \left[\mathrm{C}\sum_{\mathrm{I}\in\mathrm{D_p}} \mathrm{f}_\omega\left(\mathrm{X_{B_I}}\right)\right]\end{aligned} \tag{21}$$

where $\mathrm{D_p}$ and $\mathrm{D_n}$ are positive and negative training sets respectively. The detail solutions of CCCP algorithm for (21) are described in Algorithm 2. Finally, we can get the bag classifying rule as (17) to filter group noisy images which corresponding to unfiltered noisy query expansions.

One could expect that in our work there would be more visual patterns (responding to different query expansions) to represent the given query. In addition, MIL based methods are applied to filter group and individual noisy images to keep images from different distributions. In return the constructed image dataset could achieve better domain adaptation ability than traditional image datasets which were constructed by one query and iterative mechanism.

IV. Experiments

Since existing dataset construction methods [8], [19], [20], [35] didn't release their datasets, we can't directly compare our dataset with their extracted datasets. From another aspect,

we systematically compare the image classification ability, cross-dataset generalization ability and dataset diversity of our dataset with three publicly available datasets STL-10, CIFAR-10 and ImageNet. The motivation is domain robust image dataset should have a better image classification ability on third-party testing data. Besides, the domain robust image dataset should have a better cross-dataset generalization ability and dataset diversity. We also report the results of object detection ability of our dataset, we then compare the object detection ability of our method with four baseline methods [5], [20], [23], [31].

A. Image Dataset DRID-20 Construction

In order to facilitate comparison with datasets STL-10, CIFAR-10 and ImageNet, we choose all common categories in these datasets: airplane/aeroplane, bird, cat, dog, horse to construct our dataset. Besides, we choose other 15 categories in PASCAL VOC to construct our dataset. The reason is most of the existing weak and web supervised learning methods are tested on PASCAL VOC dataset. Overall, we use the proposed method in this paper to build a dataset DRID-20 which consists of all 20 categories in PASCAL VOC dataset.

In our experiments, for each given query (e.g.,"horse"), we first expand the given query to a set of query expansions with POS. In order to filter visual non-salient expansions, we retrieve the top $\mathbf{N} = 100$ images from image search engine as positive images (despite noisy images may be included). Set the training set and validation set $\mathrm{I_i} = \{\mathrm{I_i^t} = 75, \mathrm{I_i^v} = 25\}$, $\bar{\mathrm{I}} = \{\bar{\mathrm{I}}^{\mathrm{t}} = 25, \bar{\mathrm{I}}^{\mathrm{v}} = 25\}$. Through experiments, we declare a query expansion i to be visually salient if the classification results ($\mathrm{S_i} \geq 0.7$) giving a relatively high score. We have released the query expansions and corresponding images (original image URL) of these twenty categories on gitHub.

For less relevant expansions filtering, we select $\mathrm{n_+}$ positive training samples from these expansions which have small semantic distance or visual distance. We calculate the semantic distance and visual distance between different query (e.g., "horse" and "cow") and get the $\mathrm{n_-}$ negative training samples. We don't select the $\mathrm{n_-}$ negative training samples from these expansions which have a large semantic distance or visual distance because these expansions have a higher probability to be positive than other different query expansions. Here we set the $\mathrm{n} = 1000$ and train a classifier based on linear SVM to filter less relevant expansions.

The first $\mathbf{M} = 100$ images are retrieved from the Google image search engine for each unfiltered query expansion to construct the raw image dataset. We treat unfiltered query expansions as positive bags and images in bag as instances. We define each positive bag at least have a portion of $\delta = 0.7$ positive instances. While negative bags can be obtained by randomly sampling a few irrelevant images that are not associated with the given query. MIL based methods are applied to learn the decision function (12). Filtering individual noisy images in each bag. Besides, the decision function of (12) will also be used to select the most k positive instances in each bag, representing this bag for group noisy images filtering. For different categories, the value of k may be different. In general, categories which have larger query expansions tend to select a smaller value. In order to learn the weighting function for different distance images in the bag, as well as learn the bag classifying rule, we label 10 datasets. Each dataset contains 100 positive bags and 100 negative bags. Both of each positive bag and negative bag have 50 images. The labelling work only needs to be done once for the weighting function learning and weighted bag classifying rule (17) learning. Finally, the learned weighted bag classifying rule (17) will be used to filter noisy bags (corresponding to group noisy images).

In order to better compare with other datasets, we evenly select positive images from positive bags to construct the dataset DRID-20. Each category in DRID-20 has 1000 images. The dataset DRID-20 has been released publicly on gitHub.

TABLE I: The detail number of images for each category in related datasets

Dataset \ Category	airplane	bird	cat	dog	horse
STL-10	1300	1300	1300	1300	1300
CIFAR-10	6000	6000	6000	6000	6000
PASCAL VOC	238	330	337	421	287
ImageNet	1434	2126	1083	1603	1402
DRID-20	1000	1000	1000	1000	1000

B. General experimental set-up

For dataset image classification ability, cross-dataset generalization ability and dataset diversity comparison, we pick all five common categories in STL-10, CIFAR-10, PASCAL VOC, ImageNet and DRID-20. For object detection ability comparison, we use all 20 categories in dataset DRID-20 and PASCAL VOC.

1) General setting for image classification, cross-dataset generalization and dataset diversity: STL-10 has ten categories and each category contains 500 training images and 800 test images. All of images in STL-10 are colour 96×96 pixels. We put all training images and test images in STL-10 to represent the dataset. The CIFAR-10 dataset consists of 60000 32×32 colour images in 10 categories, with 6000 images per category including 5000 training images and 1000 test images. Similarly, we use all 6000 images for each category in CIFAR-10 to represent the dataset. ImageNet is an image dataset organized according to the WordNet [16] hierarchy. It provides on average 1000 images to illustrate each category. We use all images in ImageNet for each category to represent the ImageNet. PASCAL VOC 2007 is a benchmark dataset in image classification and object detection, providing the vision and machine learning communities with a standard dataset of images and evaluation procedures. PASCAL VOC 2007 contains 20 categories and each category has training/validation data and test data. For image classification ability, cross-dataset generalization ability and dataset diversity comparison, we utilize training/validation data to represent the PASCAL VOC 2007 dataset. Our dataset DRID-20 is constructed according to the categories in PASCAL VOC 2007 and has 1000 images for each category. To evaluate the performance of image classification ability, cross-dataset generalization ability and dataset diversity, we resize all images in STL-10, ImageNet, PASCAL VOC 2007 and our DRID-20 to 32×32. For all

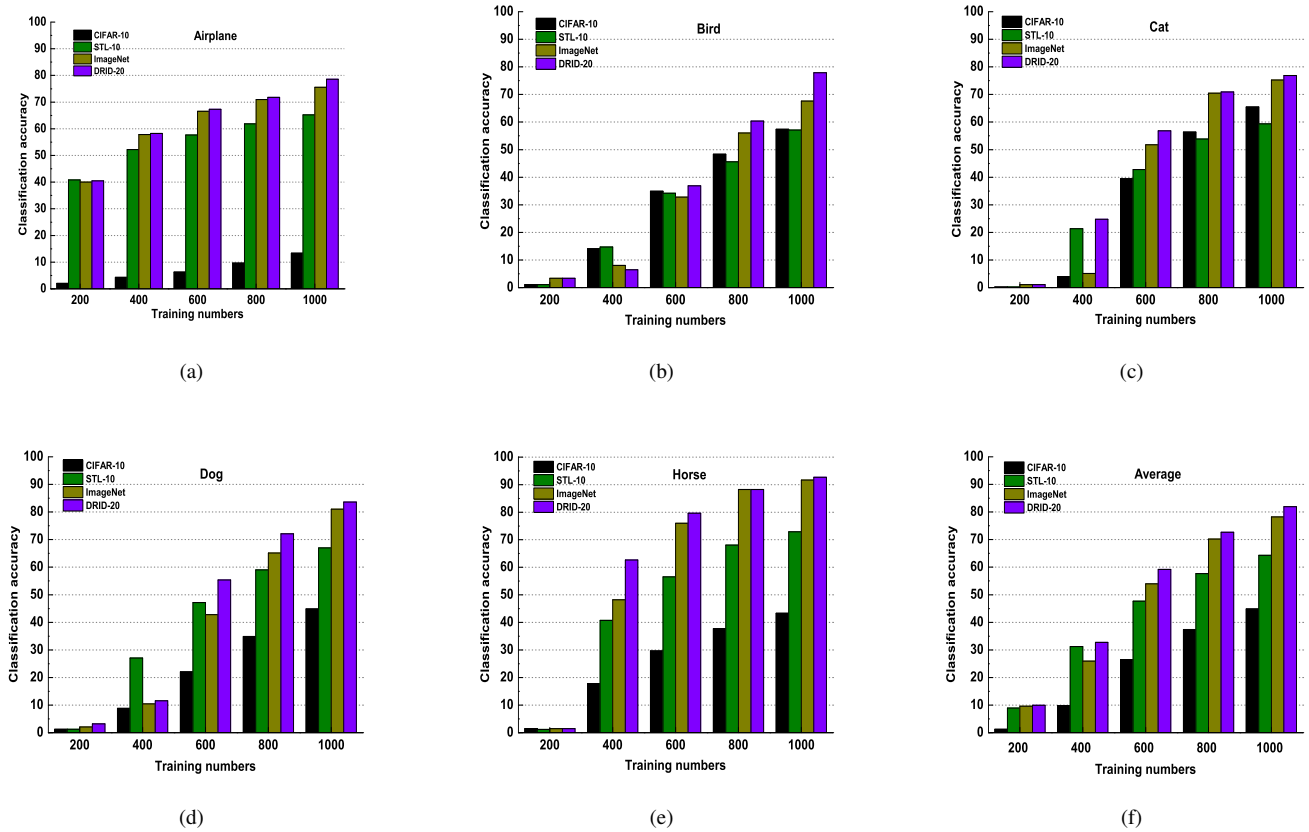

Fig. 3: The image classification ability of CIFAR-10, STL-10, ImageNet and DRID-20 on PASCAL VOC 2007 dataset: (a) airplane, (b) bird, (c) cat, (d) dog, (e) horse and (f) average.

datasets, we extract the same Histogram of Oriented Gradient (HOG) feature and train one-versus-all classifiers. The detail number of images for each category is shown in Table 1.

2) General setting for object detection: The idea of training detection models without bounding boxes has received renewed attention due to the success of the DPM [23] detector. To compare the object detection ability of our method with four other baseline methods [5], [23], [20], [31], we pick PASCAL VOC 2007 as the test data. The reason is recent state-of-the-art weakly and webly supervised methods have been evaluated on it.

For each query expansion, we train a separate DPM to constrain the visual variance. We resize images to a maximum of 500 pixels and ignore images with extreme aspect ratios (aspect ratio $>$ 2.5 or $<$ 0.4). To avoid getting stuck to the image boundary during the latent re-clustering step, we initialize our bounding box to a sub-image within the image that ignores the image boundaries. Following [23], we also initialized components using the aspect-ratio heuristic. Some of the components across different query expansion detectors end up learning the same visual pattern. For example, the images corresponding to query expansion "walking horse" are similar to the images corresponding to "standing horse". In order to select a representative subset of the components and merge similar components, we represent the space of all query expansions components by a graph $\mathrm{G} = \{\mathrm{C}, \mathrm{E}\}$, where each node represents a component and each edge represents the visual similarity between them. The score $\mathrm{d_i}$ for each node corresponds to the average precision. The weight on each edge $\mathrm{e_{i,j}}$ is obtained by running jth component detector on the ith component set. We solve for the same objective function proposed in [5] to select the representative components $\mathrm{S} \subseteq \mathrm{V}$:

$$\max_{\mathrm{S}} \sum_{\mathrm{i} \in \mathrm{V}} \mathrm{d_i} \cdot \vartheta(\mathrm{i}, \mathrm{S}) \tag{22}$$

where ϑ is a soft coverage function that implicitly pushes for diversity:

$$\vartheta(\mathrm{i}, \mathrm{S}) = \begin{cases} 1 & \mathrm{i} \in \mathrm{S} \\ 1 - \prod_{\mathrm{j} \in \mathrm{S}} (1 - \mathrm{e_{i,j}}) & \mathrm{i} \notin \mathrm{S}. \end{cases} \tag{23}$$

After the representative subset of components was obtained, we augment them with parts as described in [23] and subsequently merge all the components to produce the final detector.

C. Performance Evaluation on Image Classification Ability

In order to compare the image classification ability of our dataset DRID-20 with STL-10, CIFAR-10 and ImageNet, we choose PASCAL VOC 2007 as the third-party test data. For this experiment, we choose five categories that are common in these datasets: airplane/aeroplane, bird, cat, dog and horse.

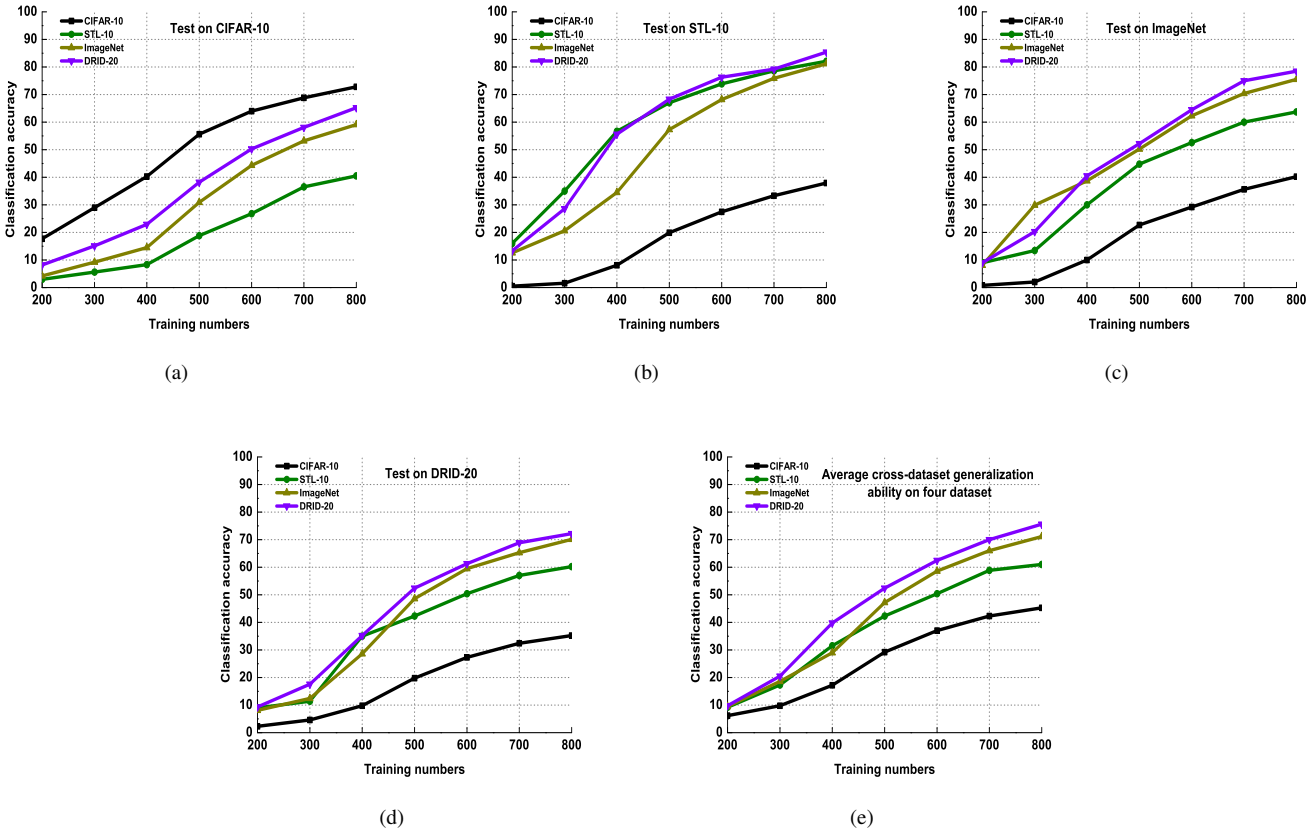

Fig. 4: Cross-dataset generalization performance of classifier learned from different datasets and then test on: (a) CIFAR-10, (b) STL-10, (c) ImageNet, (d) DRID-20, (e) Average.

For the chosen of positive training images, we randomly select training images from various datasets. For the chosen of negative training images, we fixed to use 1000 same negative training images for all datasets. For details, we sequentially select [200,400,600,800,1000] training images from CIFAR-10, STL-10, ImageNet and DRID-20 as positive training images, use 1000 fixed negative training images to learn image classifiers. Then test the performance of these classifiers on PASCAL VOC 2007 dataset (corresponding categories). We repeat the above experiment 10 times and use the average performance of image classifiers as the final performance for each dataset. The image classification ability of all datasets for each category and whole dataset is shown in Figure 3.

From Fig 3, we have the following observations:

(1) It is interesting to observe that category "airplane" has a relatively high classification accuracy than category "bird", "cat", "dog" and "horse" with a small amount of training data [200,400]. A possible explanation is that the scenes and visual patterns of "airplane" are relatively simple than category "bird", "cat", "dog" and "horse". Even with a small amount of training data, there still have a large number of positive patterns in both auxiliary and target domains. That is to say, the samples distribute densely in the feature space, the distributions of two domains are much easier overlap between each other. On the other hand, for the category "bird", "cat", "dog" and "horse", the positive samples from both domains distribute sparsely in the feature space. It is more likely that there is less overlap between the data distributions of two domains.

(2) CIFAR-10 has a much worse performance on image classification than STL-10, ImageNet and DRID-20 according to the accuracy over all 5 common categories, which demonstrates that the SVM classifier learned with training data from auxiliary domain performs poorly on the target domain. The explanation is perhaps that the data distributions of CIFAR-10 are quite different from PASCAL VOC 2007 dataset. CIFAR-10 dataset has more serious dataset bias problem than STL-10, ImageNet and DRID-20.

(3) We also observe that ImageNet is slightly worse than DRID-20 in each individual category and whole dataset, possibly because the distributions of samples from ImageNet are relatively rich. ImageNet is constructed with the goal that objects in images should have variable appearances, positions, view points, poses as well as background clutter and occlusions.

(4) DRID-20 outperforms CIFAR-10, STL-10 and ImageNet in terms of average accuracy from five common categories, which demonstrate the domain robustness of DRID-20. The explanation is DRID-20 constructed by multiple query expansions and MIL based selecting mechanisms has much more visual patterns than CIFAR-10, STL-10 and ImageNet in the condition of the same number of training samples. In other words, DRID-20 has a much richer feature distributions and it is more easily to overlap with unknown target domains.

We additionally report the hardware configuration of our experiment. For the images collection, we use two HP desktop

(a) (b)

Fig. 5: (a) Comparison of the lossless JPG file sizes of average images for five different categories in DRID-20, ImageNet and STL-10. (b) Example images from DRID-20, ImageNet, STL-10 and average images for each category indicated by (a).

PC (3.2GHz CPU with 8 Gbyte RAM). All the data processing and experiments are performed on an Acer workstation (3.5GHz CPU with 16 Gbyte RAM and 4 Gbyte VRAM) with LIBSVM [3].

D. Performance Evaluation on Cross-dataset Generalization Ability

Cross-dataset generalization ability measures the performance of classifiers learned from one dataset and tested on the other datasets. It indicates the domain robustness of dataset [21], [24], [38]. Here we compare the cross-dataset generalization ability of our dataset DRID-20 with three publicly available dataset CIFAR-10, STL-10 and ImageNet. We still choose five same categories (horse, bird, airplane, cat and dog) which are included in all four datasets to verify their cross-dataset generalization ability.

Specifically, we randomly select 200 images per category from each dataset as the test data. For the choice of training data, we sequentially select [200,300,400,500,600,700,800] images per category from various datasets as positive training images, use 1000 fixed negative training images to learn image classifiers. The training images are selected randomly in each category and training images and test images have no coincident. The average classification accuracy for five categories (horse, bird, airplane, cat and dog) represents the cross-dataset generalization ability of one dataset on another dataset. When training the image classification model, we set the same options for four datasets. Setting the type of SVM to be C-SVC, the type of kernel to be radial basis function and all other options to be the default LIBSVM options. The cross-dataset performance of four datasets and average performance for four datasets is shown in Figure 4.

By observing Figure 4, we have the following conclusions:

(1) In three of four datasets, with the increase of training images, the best performance of classification is achieved by DRID-20. When tested on STL-10, ImageNet and DRID-20, it shows that the generalization ability of our dataset DRID-20 is very close and DRID-20 performs slightly better than STL-10 and ImageNet. Besides, DRID-20 outperforms CIFAR-10, STL-10 and ImageNet in terms of average cross-dataset performance on four datasets, which demonstrate the domain robustness of DRID-20. A possible explanation is our dataset DRID-20 constructed by multiple query expansions has much more visual patterns or feature distributions than STL-10, CIFAR-10 which just using one query for candidate images collection. At the same time, MIL based selecting mechanisms maximize the retention of the useful visual patterns to represent the dataset DRID-20.

(2) CIFAR-10 shows a poor performance on cross-dataset generalization except on its own dataset. The explanation is that the data distributions of its auxiliary domain and target domain are quite related, making it difficult for other datasets to exceed its performance when tested on CIFAR-10. All images in CIFAR-10 are cut to 32×32 and objects in these images are located in the middle of the image. Besides, these images contain relatively small other objects or scenes. Images in STL-10 are 96×96 and in ImageNet and DRID-20 are full size. These images not only contain target objects, but also include a large number of other scenarios or objects. Based on these conditions, although CIFAR-10 has a better performance on its own domain, it still has a serious dataset bias problem which coincide with its average cross-dataset generalization performance.

E. Performance Evaluation on Dataset Diversity

Following [1], [29], we compute the average image of each category and measure lossless JPG file size which reflects the amount of information in an image. The basic idea is that diverse image dataset will result in a blurrier average image, whereas an image dataset with little diversity will result in a more structured, sharper average image. Therefore, we expect to see a smaller JPG file size of the average image of a more diverse image dataset.

TABLE II: Object detection results (A.P.) on PASCAL VOC 2007 (test).

Method	[20]	[31]	[5]	**Our**	[23]
Supervision	weak	weak	web	web	full
airplane	13.4	**17.4**	14.0	15.5	33.2
bike	**44.0**	-	36.2	40.6	59.0
bird	3.1	9.3	12.5	**16.1**	10.3
boat	3.1	9.2	**10.3**	9.69	15.7
bottle	0.0	-	9.2	**13.7**	26.6
bus	31.2	-	35.0	**42.0**	52.0
car	**43.9**	35.7	35.9	37.9	53.7
cat	7.1	9.4	8.4	**9.8**	22.5
chair	0.1	-	**10.0**	9.6	20.2
cow	9.3	9.7	17.5	**18.4**	24.3
table	9.9	-	6.5	**10.6**	26.9
dog	1.5	3.3	**12.9**	11.6	12.6
horse	29.4	16.2	30.6	**36.1**	56.5
motorcycle	**38.3**	27.3	27.5	36.9	48.5
person	4.6	-	6.0	**7.9**	43.3
plant	0.1	-	**1.5**	1.3	13.4
sheep	0.4	-	18.8	**20.4**	20.9
sofa	3.8	-	10.3	**10.8**	35.9
train	**34.2**	15.0	23.5	27.6	45.2
tv/monitor	0.0	-	16.4	**18.4**	42.1
average	13.87	15.25	17.15	**19.74**	33.14

We resize all images in STL-10, ImageNet and DRID-20 to 32×32 gray images, and create average images for each category from 100 randomly sampled images. Fig. 5 compares the image diversity of five common categories in DRID-20, ImageNet, STL-10 and shows example images and average images in these datasets. By observing Fig. 5(a), the average image of DRID-20 is blurred and hard to recognize out the object, while the average image of ImageNet and STL-10 is relatively more structured and sharper. DRID-20 has slightly smaller JPG file size than ImageNet and STL-10. This phenomenon is universal for five same categories.

DRID-20 is constructed with the goal that images in this dataset should exhibit domain robustness and can effectively alleviate the dataset bias problem. In order to obtain domain robustness, we not only consider the source of candidate images, but also retain the images from different distributions. Through above experiments, with a certain number of samples, DRID-20 contains much more effective visual patterns and feature distributions than dataset CIFAR-10, STL-10 and ImageNet. Thus it presents a better domain adaptation ability.

F. Performance Evaluation on Object Detection Ability

We report the performance of object detection on PASCAL VOC 2007 test set. Table 2 shows the results of our proposed method and compares to the state-of-the-art weakly and webly supervised methods [5], [20], [31]. Method [20] and [31] have state-of-the-art performance for weakly-supervised object detection. [31] trains on manually selected videos without bounding boxes and shows results on 10/20 categories. [20] uses weak human supervision (VOC data with image-level labels for training) and initialization from objectness [26]. [5] takes web supervision and then trains a mixture DPM detector for the object. [23] is currently fully supervised object detection method and it is a possible upper bound for weakly and webly supervised approaches.

Compared to [20], [31] which uses weak supervision and [23] which uses full supervision, the training set of our proposed approach and [5] do not need to be labelled manually. Nonetheless, the results of our proposed approach and [5] surpass the previous best results in weakly supervised object detection method [20], [31]. A possible explanation is perhaps that both of our approach and [5] use multiple query expansions for candidate images collection, the training data collected by our approach and [5] are more rich and contains more effective visual patterns. Compared to [5] which also uses web supervision and multiple query expansions for candidate images collection, our method surpasses their results in most of the cases. The explanation is we take different mechanisms for noisy images removing. Compared to [5] which takes iterative mechanisms in the process of noisy images filtering, our approach applies MIL based method for noisy images removing. It can maximize keeping images from different data distribution while with noisy images filtered out.

By using the same feature and training strategies, our approach achieves a better performance than weakly and webly supervised method [20], [31], [5]. The main reason for this is that our training data generated from multiple query expansions and MIL based filtering mechanisms contains much richer and accurate visual descriptions for these categories. In other words, our approach discovers much more useful linkages to the visual patterns for the given category.

V. Conclusion and Future Work

In this work, we presented a new framework for domain robust image dataset construction with web images. Three successive modules were employed in the framework including query expanding, noisy expansions filtering and noisy images filtering. To verify the effectiveness of our proposed method, we constructed an image dataset with 20 categories. Three extensive experiments showed that our framework had a better domain adaptation ability than traditional datasets like STL-10, CIFAR-10 and ImageNet which were constructed by one query. Besides, our approach had a better object detection ability and can surpass [5], [20], [31] in most of the cases. The reason is our dataset has much more visual patterns (responding to different query expansions) to represent the given query, so our dataset can be adapted to unseen domains which may have different visual feature distributions.

Although good results were obtained, there is still room to improve the proposed dataset construction framework. For example, we can potentially use more sophisticated approaches to purify noisy query expansions, noisy images and that will be the focus of our future work.

Acknowledgments

This research was supported by the Natural Science Foundation of China (No. 61473154).

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[2] F. R. Bach, G. R. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the International conference on Machine learning*, pages 220–228. ACM, 2004.

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

[4] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[5] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277. IEEE, 2014.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[8] X.-S. Hua and J. Li. Prajna: Towards recognizing whatever you want from images without image labeling. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 137–144. AAAI, 2015.

[9] J. E. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

[10] S.-J. Kim and S. Boyd. A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM Journal on Optimization*, 19(3):1344–1367, 2008.

[11] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, Citeseer, 2009.

[12] T. Leung, Y. Song, and J. Zhang. Handling label noise in video classification via multiple instance learning. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2056–2063. IEEE, 2011.

[13] L.-J. Li and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *International journal of computer vision*, 88(2):147–168, 2010.

[14] Y.-F. Li, I. W. Tsang, J. T. Kwok, and Z.-H. Zhou. Tighter and convex maximum margin clustering. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 344–351, 2009.

[15] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174. Association for Computational Linguistics, 2012.

[16] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[17] L. Niu, W. Li, and D. Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2774–2783, IEEE, 2015.

[18] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[19] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 33(4):754–766, 2011.

[20] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *Proceedings of IEEE International Conference on Computer Vision*, pages 343–350. IEEE, 2011.

[21] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011.

[22] S. Vijayanarasimhan and K. Grauman. Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[23] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32 (9) (2010) 1627–1645.

[24] Y.-Z. Yao, J. Zhang, F.-M. Shen, X.-S. Hua, J.-S. Xu, and Z.-M. Tang. Automatic image dataset construction with multiple textual metadata. In *Proceedings of IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2016.

[25] A. Coates, A. Y. Ng, H. Lee, An analysis of single-layer networks in unsupervised feature learning, in: International conference on artificial intelligence and statistics, 2011, pp. 215–223.

[26] B. Alexe, T. Deselaers, V. Ferrari, Measuring the objectness of image windows, Pattern Analysis and Machine Intelligence, IEEE Transactions on 34 (11) (2012) 2189–2202.

[27] Lixin Duan, Wen Li, Ivor Wai-Hung Tsang, and Dong Xu, "Improving web image search by bag-based reranking," *Image Processing, IEEE Transactions on*, vol. 20, no. 11, pp. 3280–3290, 2011.

[28] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset.

[29] B. Collins, J. Deng, K. Li, L. Fei-Fei, Towards scalable dataset construction: An active learning approach, in: Computer Vision–ECCV 2008, Springer, 2008, pp. 86–98.

[30] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 3485–3492.

[31] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari, "Learning object class detectors from weakly annotated video," in *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*. IEEE, 2012, pp. 3282–3289.

[32] B. Siddiquie, A. Gupta, Beyond active noun tagging: Modeling contextual interactions for multi-class active learning, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 2979–2986.

[33] S. Vijayanarasimhan, K. Grauman, Large-scale live active learning: Training object detectors with crawled data and crowds, International Journal of Computer Vision 108 (1-2) (2014) 97–114.

[34] K. C Kiwiel, "Proximity control in bundle methods for convex nondifferentiable minimization," *Mathematical programming*, 46(1-3) (1990) 105–122.

[35] L.-J. Li, L. Fei-Fei, Optimol: automatic online picture collection via incremental model learning, International journal of computer vision 88 (2) (2010) 147–168.

[36] R. Speer, C. Havasi, Conceptnet 5: A large semantic network for relational knowledge, in: The Peoples Web Meets NLP, Springer, 2013, pp. 161–176.

[37] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 160–167.

[38] Y.-Z. Yao, X.-S. Hua, F.-M. Shen, J. Zhang, and Z.-M. Tang. A domain robust approach for image dataset construction. In *Proceedings of ACM International Conference on Multimedia*, ACM, 2016 accepted.

[39] Gustavo Carneiro, Antoni B Chan, Pedro J Moreno, and Nuno Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[40] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[41] Alan L Yuille and Anand Rangarajan, "The concave-convex procedure," *Neural computation*, vol. 15, no. 4, pp. 915–936, 2003.