Exploiting Web Images for Weakly Supervised Object Detection

Qingyi Tao, Hao Yang, and Jianfei Cai, *Senior Member, IEEE*

***Abstract*—In recent years, the performance of object detection has advanced significantly with the evolution of deep convolutional neural networks. However, the state-of-the-art object detection methods still rely on accurate bounding box annotations that require extensive human labeling. Object detection without bounding box annotations, that is, weakly supervised detection methods, are still lagging far behind. As weakly supervised detection only uses image level labels and does not require the ground truth of bounding box location and label of each object in an image, it is generally very difficult to distill knowledge of the actual appearances of objects. Inspired by curriculum learning, this paper proposes an easy-to-hard knowledge transfer scheme that incorporates easy web images to provide prior knowledge of object appearance as a good starting point. While exploiting large-scale free web imagery, we introduce a sophisticated labor-free method to construct a web dataset with good diversity in object appearance. After that, semantic relevance and distribution relevance are introduced and utilized in the proposed curriculum training scheme. Our end-to-end learning with the constructed web data achieves remarkable improvement across most object classes, especially for the classes that are often considered hard in other works.**

***Index Terms*—Weakly supervised learning, object detection, curriculum learning.**

I. Introduction

With the rapid growth of computational power and dataset size and the development of deep learning algorithms, object detection, one of the core problems in computer vision, has achieved promising results [18], [19], [21], [23]. However, state-of-the-art object detection methods still require bounding box annotations which cost extensive human labour. To alleviate this problem, weakly supervised object detection approaches [2]–[4], [6], [7], [12], [16], [17], [25], [27], [29] attracted much attentions. These approaches aim at learning an effective detector with only image level labels, so that no labour-extensive bounding box annotations are needed. Nevertheless, as objects in common images can appear in different sizes and locations, only making use of image level labels are often not specific enough to learn good object detectors, and thus the performance of most weakly supervised methods are still subpar compared to their strongly supervised counterparts, especially for small objects with occlusions, such as "bottle" or "potted plant". As shown in Fig. 1, images containing small objects or with very complicated contexts are hard to learn. In contrast, images containing a single object with very clean background provides very good appearance priors for learning object detectors. Particularly, for these easy images, the difficulty of localizing the objects is much lower than complicated images. With correct localization, the appearance model can be better learned. Therefore, easy images can provide useful information about object appearance for learning the model for more complicated images. Unfortunately, such easy images are rarely available in object detection datasets, such as PASCAL VOC or MS COCO, as images in these multi-object datasets usually contain cluttered objects and very complicated background. On the other hand, there are a large number of easy web images available online and we can exploit these web images for the weakly supervised detection (WSD) task.

Fig. 1. Easy web images and VOC images. Web images have clean background while VOC images are more difficult with cluttered instances and complicated background.

However, to construct a suitable auxiliary dataset and appropriately design an algorithm to utilize the knowledge from the dataset are non-trivial tasks. In this paper, we intend to provide a practical and effective solution to solve both problems.

Manuscript received August 3, 2017; revised April 15, 2018 and July 29, 2018; accepted September 20, 2018. Date of publication October 11, 2018; date of current version April 23, 2019. This work was supported in part by MoE Tier-2 under Grant 2016-T2-2-065, in part by Tier-1 under Grant 2018-T1-001-115, and in part by NTU CoE under Grant 2016. This work was done when Hao Yang was at Nanyang Technological University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Crandall. *(Corresponding author: Qingyi Tao.)*

Q. Tao is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 and also with the NVIDIA AI Technology Center, Singapore 138522 (e-mail: qtao002@e.ntu.edu.sg).

H. Yang is with Amazon, Seattle, WA 98133, USA (e-mail: lancelot365@gmail.com).

J. Cai is with the School of Computer Science and Engineering, , Nanyang Technological University, Singapore 639798 (e-mail: asjfcai@ntu.edu.sg).

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

Digital Object Identifier 10.1109/TMM.2018.2875597

Specifically, various image search engines like Bing, Google, Flickr provide access to freely available web data of high quality images. Recent researches [5], [9], [13], [20], [31], [32] have already utilized these large-scale web data in various vision tasks. However, as object detection tasks impose specific requirements for auxiliary web data, we need to carefully design a labour-free way to obtain suitable images for the task.

First of all, when constructing the web dataset, we need to consider the relevance of web images in order to effectively transfer the knowledge of easy web images to the target detection dataset. In this paper, we break down this relevance into two parts, namely semantic relevance, which refers to the relevance between web images and the target labels, and the distribution relevance, which refers to the relevance between web images and target images. As we will show in later sections, the semantic relevance focuses on a larger picture in the semantic space, while the distribution relevance measures more fine-grain differences in the feature distributions. To give an example, for category "chair", the semantic relevance measures whether a certain web image is "chair" or not, and the distribution relevance measures whether this web image lies on the manifold formed by the specific "chairs" in the target dataset.

Secondly, apart from the relevance problem, we also need to consider the diversity of the web images. As sub-categories, poses as well as backgrounds are crucial for the success of object detection, our web images should not only be easy and related to the target dataset, but also contain a variety of different images even for the same category. With single text query, commonly used image search engines are not able to produce images with large intra-category diversity, especially in top ranked results. Therefore, inspired by [9], which uses ngrams to retrieve the fine-grained dataset, and [33], which expands query words to construct a domain robust dataset, we propose a multi-attribute web data generation scheme to enhance the diversity of web data. Specifically, we construct a general attribute table with common attributes that can easily be propagated to other target datasets as well. With the attribute table, we are able to build a hassle-free web dataset with proper category-wise diversity for the coarsely labeled dataset.

Once we have an appropriate web dataset, we need to consider how to transfer the knowledge from the easy web images to more complex multi-object target datasets. During the recent years, easy web images have been used in other weakly supervised tasks, such as weakly supervised segmentation [30]. To the best of our knowledge, we are the first work bringing in web images for improving the weakly supervised object detection task.

Inspired by curriculum learning [1], we propose a simple but effective hierarchical curriculum learning scheme. Specifically, with the hierarchical curriculum structure, all web images are considered easier than target images, which we refer to as the first level of curriculum, followed by the second level of curriculum that includes all target images. Extensive experimental results show that our constructed web image dataset and the adopted curriculum learning can significantly improve the WSD performance.

Our paper is organized as follows. In Section II, we review the related works. In Section III, we describe our methodology on constructing web image dataset and hierarchical curriculum learning. In Section IV, we evaluate our method on two widely used benchmark datasets. Lastly, we conclude the paper in Section V.

II. Related Work

Our work is related to several areas in computer vision and machine learning.

Weakly Supervised Object Detection (WSD): Traditional WSD methods like [6] address this problem with multiple instance learning (MIL) [8], which treats each image as a bag and each proposal/window in the image as an instance in the bag. A positive image contains at least one positive instance whereas a negative image contains only negative instances. Since MIL approaches alternate the processes between selecting a region of objects and using the selected region to learn the object appearance model, they are often sensitive to initialization and often get stuck in local optima. [4] proposed a two-stream CNN structure named WSDDN to learn localization and recognition in dedicated streams respectively. These two streams share the common features from the earlier convolutional layers and one fully connected layer. It learns one detection stream to find the high responsive windows and one recognition stream to learn the appearance of the objects. In this way, the localization and recognition processes are decoupled. Similarly, [12] also uses a two-stream structure and additionally involves the contextual feature in the localization stream. Another direction of WSD has been proposed in [25] to utilize deformable part-based models for WSD problem. Recently, [7], [17] incorporate segmentation into weakly supervised detection to improve the detection performance. [24] introduces a multi-instance classification refinement network to the WSDDN [4] for simultaneously refining the localization of objects and significantly improves the results.

In this research, we mainly use WSDDN [4] as an example to evaluate our learning method. Since WSDDN separates recognition and localization into two individual streams, it introduces additional degrees of freedom while optimizing the model, and hence it is hard to train at the early stage. It is also sensitive to initialization. Thus, in this work we propose to explicitly provide good initialization during the training process in an easy-to-hard manner. We further experiment our method on top of [24] and show that our learning scheme is a general approach that can be applied to improve other WSD methods as well.

Exploiting Additional Data: Our work is also related to studies on exploiting additional data for weakly supervised learning [30], [14]. [30] uses web images to create pseudo ground truth for weakly supervised semantic segmentation task and improves the performance significantly. Later, for object detection task, [26] uses the same web data to learn detectors by transferring web knowledge. However, this work focuses on the "zero-annoation" setting without using any image-level labels in target dataset, the performance is not comparable with weakly supervised detection approaches. [14] also uses additional data for pseudo ground truth creation. However, instead of using free web images, [14] tracks objects in videos to create pseudo ground truth and uses the generated pseudo ground truth

to train a fully supervised object detector. In [14], it uses a public video dataset named "YouTube-Objects" with video level annotations, which is not free of human annotation and not extensible for more categories. There are only 10 overlapped classes between "YouTube-Objects" and Pascal VOC, and therefore only 10 classes are evaluated in [14]. In our work, we can freely exploit the accessible web images and have the flexibility to make use of different properties of web data for our task.

Curriculum Learning: Our work is inspired by curriculum learning [1] scheme. Curriculum learning was initially proposed to solve the shape recognition problem, where the recognition model is first trained to recognize the basic shapes and then trained on more complicated geoshapes. Recently, Tudor *et al.* [28] used this easy-to-hard learning scheme in MIL problem but mainly focused on learning a model to rank images with difficulty that matches the human perspective. In our work, we propose a hierarchical curriculum scheme that incorporates easy web images in early training stage to provide prior knowledge for the subsequent training on complicated images.

Learning from Weak or Noisy Labels: This paper is also related to those works on learning from weak or noisy labels [5], [9], [10], [22], [33]. In [9], they proposed a classifier-based cleaning process to deal with the noisy labels. They first train a classification model on images with higher confidence and then use this model to filter the outliers in the rest of images. Later, with the incorporation of CNN, a novel loss layer is introduced to the deep network in [22]. In [5], web images are separated into easy images (Google) and hard images (Flickr). They build a knowledge graph on easy web images and use the graph as a semantic constraint to deal with the possible label-flip noises during training of harder web images. Similarly, [10] learns the mutual relationship to suppress the feedback of noises during back propagation. These works emphasize their methods to lessen the impact by outliers during the training process. In our work, apart from the outliers, we also consider the distribution mismatch problem since we acquire web data that are from completely different information sources with discrepant distribution compared to the target dataset.

III. Approach

In this part, we will first briefly introduce the base model WSDDN [4] since we will use this state-of-the-art weakly supervised objection algorithm as an example to show the effectiveness of our scheme. Then we introduce the methodology on constructing the web dataset and the hierarchical curriculum learning to transfer the knowledge of web images to the target dataset. Note that our scheme is general and can also be adapted to any other available algorithms if necessary.

A. WSDDN

We first introduce weakly supervised deep detection network, or WSDDN [4], which is utilized as a baseline for our experiments. WSDDN provides an end-to-end solution that breaks the cycle of training of classification and localization alternatively by decoupling them into two separate streams.

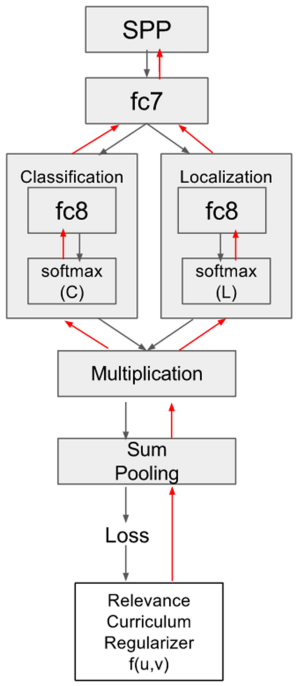

Fig. 2. WSDDN with relevance curriculum regularizer. The relevance curriculum regularizer suppresses backpropagation from samples which do not fit in the relevance region and curriculum region.

Particularly, WSDDN replaces the last pooling layer with spatial pyramid pooling layer [15] to obtain SPP features of each region of interest (RoI). As shown in Fig. 2, the SPP features are passed to a classification stream and a localization stream which individually learns the appearance and location of the objects. In the classification stream, the score for each RoI from $fc8$ layer is normalized across classes to find the correct label of RoIs. In the localization stream, the scores of all RoIs are normalized category-wise to find most respondent RoIs for each category. Then the probability outputs from both softmax layers are multiplied as the final detection scores for each RoI. Finally, detection scores of all RoIs are summed up to one vector as the image level score to optimize the loss function (1).

$$L(y_{ci}, x_i|w) = -log\left(y_{ci}\left(\Phi_c(x_i|w) - \frac{1}{2}\right) + \frac{1}{2}\right) \quad (1)$$

In the binary log loss function $L(y_{ci}, x_i|w)$, x_i is the input image i, and y is the binary image level label where $y_{ci} = \{-1, 1\}$ for class c in image i. Output from the last sum pooling layer is denoted as $\Phi_c^y(x_i|w)$ which is a vector in range of 0 to 1 with the dimension equal to the number of categories. For each class c, if the label y_{ci} is 1, $L(y_{ci}, x_i|w) = -log(p(y_{ci} = 1))$ and if y_{ci} is -1, $L(y_{ci}, x_i|w) = -log(1 - p(y_{ci} = 1))$.

This network learns quite slowly at the first few epochs for complicated images because both detection and recognition streams are randomly initialized without any prior information. However, learning simple images with a single large object with a clean background is much easier. The network can easily locate the object and learn the common appearance based on the location. Therefore, we intend to use easy web images to learn the appearance first and transfer the model to learn hard images.

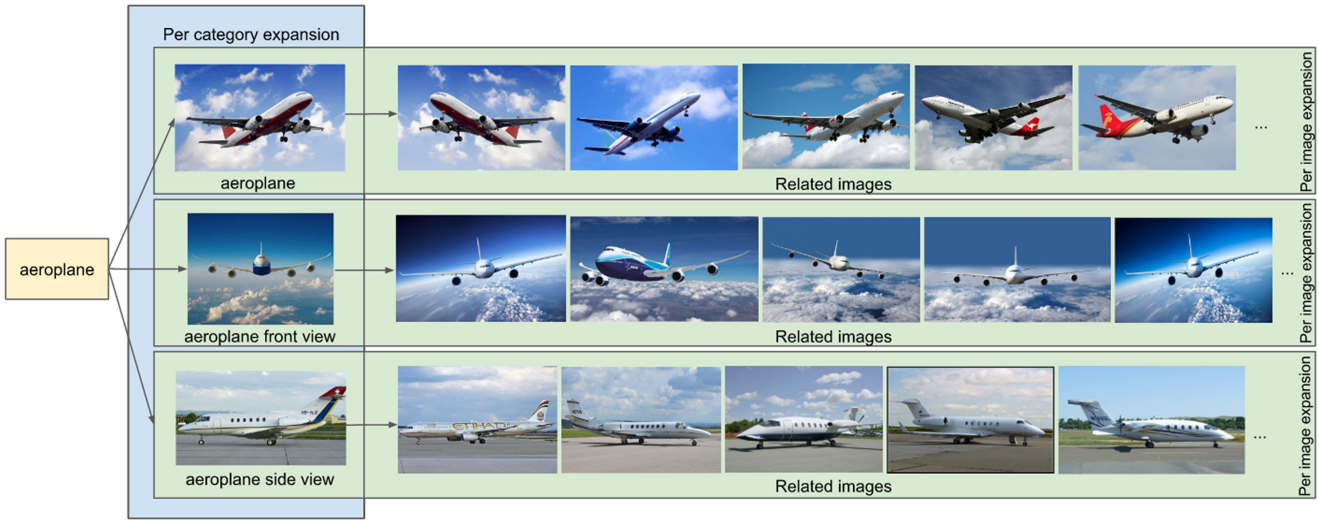

Fig. 3. Multi-attribute related dataset. Aeroplane category is expanded with multi-view attributes including front view and side view. Each multi-attribute web image is then expanded by the related images obtained from Bing image search engine.

B. Constructing Multi-Attribute Web Dataset

In this section, we describe our method to construct a diversified and robust web dataset by introducing an expand-to-condense process. Specifically, we first introduce multiple attributes on top of the given target labels when crawling for web images to improve the generalization ability of the obtained dataset. Then we introduce both semantic relevance and distribution relevance to condense the dataset by filtering out irrelevant images.

1) Expand to Diversify: Free web images are abundantly available and accessible. Many image search engines can provide high quality images by searching for the object names, such as Google, Flickr and Bing. In our preliminary study, we observe that images retrieved by Bing are generally easier than images from other search engines. Since easier images are intuitively better for learning object appearance, we choose Bing as the search engine to crawl web images.

However, for most search engines, we observed that if we just use the given target labels as keywords, the resulting images are very similar in object appearances, poses or sub-categories. Moreover, the number of good quality images returned per query is very limited and lower ranked images are generally very noisy and unrelated to the queries.

To solve the problem of lacking diversity as well as limited number of high quality images, we introduce multiple attributes to each category. Based on the general knowledge of object detection, we define a set of attributes in three general aspects: namely viewpoints, poses or habitats of the objects.

First of all, adding viewpoint attributes such as "front view" and "side view" not only provides extensive amount of high quality images for artificial objects like "aeroplane", "car" and "bus", but also enhances the appearance knowledge of these objects, which will eventually make the detector more robust. Note that for categories without clear discrepancy between front view and side view such as "bottle" and "potted plant", as well as flat objects like "tv monitor", we do not include these attributes. Secondly, for animals like "cat" and "dog", we add pose attributes. As their appearances vary significantly in different poses, adding such attributes will also be beneficial towards building a more robust detector. In particular, we add poses such as "sitting", "jumping" and "walking" to these animal categories. Last but not least, for category "bird" which resides in different habitats, we add habitat attributes of "sky" and "water". The set of attributes is summarized in Table I. Note that following the same spirit, the table can be easily expanded to other categories.

TABLE I
ATTRIBUTE TABLE

Category	Viewpoint	Pose	Habitat
aeroplane; bicycle; boat; bus; car; motorbike; train; chair; diningtable; sofa	front view; side view	–	–
bird	front view; side view	–	water; sky
cat; dog	front view; side view	sitting; walking; jumping	–
cow; sheep	front view; side view	walking;	–
horse	front view; side view	walking; jumping	–
person	front view; side view	sitting; standing; walking	–

Moreover, to overcome the limitation of few clean images available in the top ranking, we also crawl related images. Related images are the images retrieved with similar visual appearance by using each of the previously retrieved top ranked images as a query to the search engine. These related image can expand the size of the web dataset by more than 20 times and also introduce more variations to the dataset. Fig. 3 illustrates the process of expanding the dataset by

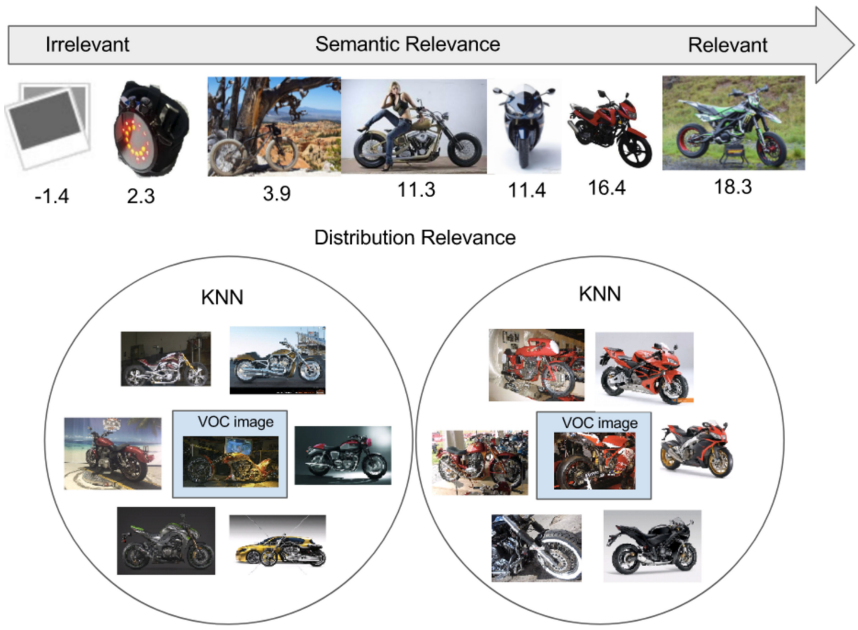

Fig. 4. Illustration of relevance metrics including semantic relevance and distribution relevance. Top: semantic relevance by the scores from web-to-web classifier for motorbike images, where non-meaningful images have negative scores, outliers with wrong objects have very low scores, and images containing correct objects have high scores. Bottom: distribution relevance by k nearest neighbors of each motorbike image in VOC dataset, where images in the neighborhood of VOC images with small feature distances are considered relevant to the target dataset.

the multi-attribute per-category expansion and the per-image expansion.

2) Condense to Transfer: Once we obtain a large scale web image dataset, we are facing with the relevance problem. As free web data often contain many noisy images, to effectively make use of these web images, we need to analyse the image relevance to condense the noisy data. In this paper, we break down the image relevance to two parts: semantic relevance and distribution relevance. In detail, semantic relevance indicates whether a image contains the correct objects and distribution relevance measures how well a web image matches the distribution of the target dataset.

Firstly, to measure the semantic relevance, we train a web-to-web outlier detector to find images with wrong labels in the web dataset. Specifically, we select top 80 images from queries of each target label and top 20 images from queries of each attribute + label combination. As we only use high ranked images as seed images, the "cleanness" of the images can be guaranteed, and thus we are able to learn a more robust outlier detector.

The outlier detector is trained iteratively with the expansion of the seed images. Similar to the idea of active learning, we train a CNN classifer with softmax loss with the seed images. Then it is applied to the whole set of web images. The highly confident positive samples are then used as the second batch of training images for the next iteration. After a few iterations, the classification scores from the final stabilized model are used to measure semantic relevance. As shown in Fig. 4, our model can provide a very solid semantic relevance measurement. Most of the non-meaningful images have negative scores, outliers with wrong objects have very low scores and images with correct objects have high scores.

Secondly, since semantic relevance condenses images purely based on their semantic meaning regardless of the distribution matching with the target dataset, we also consider the distribution relevance for more fine-grain measurements. To align the diversified web dataset into the distribution of target dataset, we search in the neighborhood of the target dataset to find similar web images. Particularly, for each single-label image in the target dataset, we select k nearest web images in the feature space. The distance between images is defined as the Euclidean distance between their corresponding CNN features. Specifically, we use the L2 normalized $fc7$ feature from a pretrained vgg-f model with PCA dimension reduction to represent each image. As shown in Fig. 4, our method is capable of extending the target dataset with web images having very similar object appearances and poses.

We expect both relevance metrics to be effective for this task since it is intuitive to eliminate noises and unrelated data during the training. Nevertheless, our experiment result shows that matching the web data to target distribution is not as helpful as using a clean but diversified web dataset.

C. Relevance Curriculum Regularizer

Incorporating a good quality web dataset to the target dataset does not automatically guarantee better performance. Based on our experiments, we find out that simply appending these web images to target dataset is unhelpful or even harmful. These easy web images could lead to skewed training models due to the distribution misalignment problem of the two datasets.

Therefore, instead of simply appending web data to target dataset, we propose a hierarchical curriculum structure. Specif-

ically, we first consider a coarser curriculum with web images as easy and all target images as hard. If necessary, we could also add a fine curriculum to each dataset for full curriculum learning. Moreover, in addition to the normal curriculum or self-paced learning [11], we also consider adding an extra relevance term. As an analogy, we could consider web images as extracurricular activities. In order to help students with their learning, extracurricular activities need to be relevant to the course, in the same way that we should learn from easy images and relevant images.

In particular, to incorporate both curriculum and relevance constraints in training, we propose a relevance curriculum regularizer to the base detection structure:

$$E(w) = \sum_{i=1}^{n} \sum_{c=1}^{C} L(y_i, x_i|w) \cdot f(u_i, v_i), \tag{2}$$

$$f(u_i, v_i) = \sigma(u_i) \cdot \psi(v_i), \tag{3}$$

where u_i is the relevance variable indicating whether the training sample is relevant as discussed in III-B. v_i is the curriculum regulation variable which indicates the difficulty score of each image. σ is the relevance region function used so that only relevant samples can be learned every epoch. If a sample is in the relevance region, the value of $\sigma(u)$ is 1 and otherwise 0. ψ is the curriculum region. It controls the pace of learning that allows only easy samples to be learned at early stage and gradually adding harder samples along the training process. If the difficulty score of sample image is within the curriculum region, $\psi(v)$ is 1 and otherwise, $\psi(v)$ is 0. As described previously, we implemented a hierarchical curriculum, where $\psi(v)$ for all web images are consider as 1 first, then we gradually expand it to include target images.

IV. Experiments

In this section, we evaluate the effectiveness of our proposed weakly supervised object detection.

A. Model Setting & Datasets

Similar to the original WSDDN work, our method requires a proposal method to generate bounding box proposals. For a fair comparison, same as in the original WSDDN, we use Edgebox [34] as the proposal method to generate around 2000 bounding boxes. In fact, since the web images are simple images with a clean background, only hundreds of proposals are generated with Edgebox. For the more complicated target images, following the original setting, we select up to 2000 proposals with high Edgebox objectness scores. To train the network, we use the vgg-f model pretrained on ImageNet as the initial model. For fairness, our results are compared with the baseline method trained on vgg-f as well.

We evaluate our method on PASCAL VOC2007 and VOC2012 datasets with 20 object categories. During the training, we use only image-level labels of the training images. The evaluation metric is the commonly used detection mAP with IoU threshold of 0.5.

TABLE II
Results of the Detection mAP on VOC2007 Test Set by Using the Curriculum Regularization Term to Train VOC2007 Trainval Set

Methods	mAP
WSDDN (baseline)	33.9
CurrWSDDN	**35.5**

Our multi-attribute related web dataset has been released for future studies.[1] Apart from our multi-attribute related web images from Bing, our experiments also evaluate the performance using STC Flickr clean dataset [30]. STC Flickr clean dataset is used as additinal data for weakly supervised semantic segmentation task in [30] and contributes significant improvements.

B. Results Regarding Curriculum Learning

We first evaluate the effectiveness of applying the curriculum learning method on PASCAL VOC2007 trainval set itself, without using our web data. This is to verify our scheme of easy-to-hard learning for a weakly supervised detector. The curriculum is designed by the ranking of the mean edge strength of each image. The mean edge strength of an image is defined as the number of edge pixels over the total number of pixels. This is a simple yet intuitive method because images with more edges tend to have a more complicated background or contain more cluttered objects, and thus it is reasonable to consider them as hard samples. Fig. 5 gives some examples, which show that the mean edge length represents the relative difficulty of the images well.

Specifically, we use the classical LoG edge detector to detect edges for simplicity. We also experimented with Canny Edge detector that generates similar edge strength results. The purpose of using edge detectors is to compute the mean edge strengths of images for relative ranking. Thus, any edge detector can be applied here. For each curriculum region, we add $\frac{1}{5}$ of more difficult images from each category. This is to balance the number of positive samples from each category in every iteration. In this way, the curriculum consists of five overlapped regions with gradually increased image complexity. Table II shows the detection result ('CurrWSDDN') of applying the curriculum regularization term to train VOC2007 trainval set only, compared with the result of the baseline ('WSDDN'). We can see that using curriculum learning on VOC2007 training images alone already improves the performance. This suggests that for training weakly supervised object detector, it is beneficial to train the network in an easy-to-hard manner. Note that the baseline WSDDN result is obtained by running the original WSDDN codes released in Github with the same setting,[2] which is slightly different from the result of 34.5 reported in [4].

C. Results Regarding Constructed Web Dataset

Having proven that curriculum learning is helpful on weakly supervised object detection, we now evaluate the usefulness of

[1]Our dataset is available at https://github.com/truetqy/Exploiting-Web-Images-for-Weakly-Supervised-Object-Detection

[2]https://github.com/hbilen/WSDDN

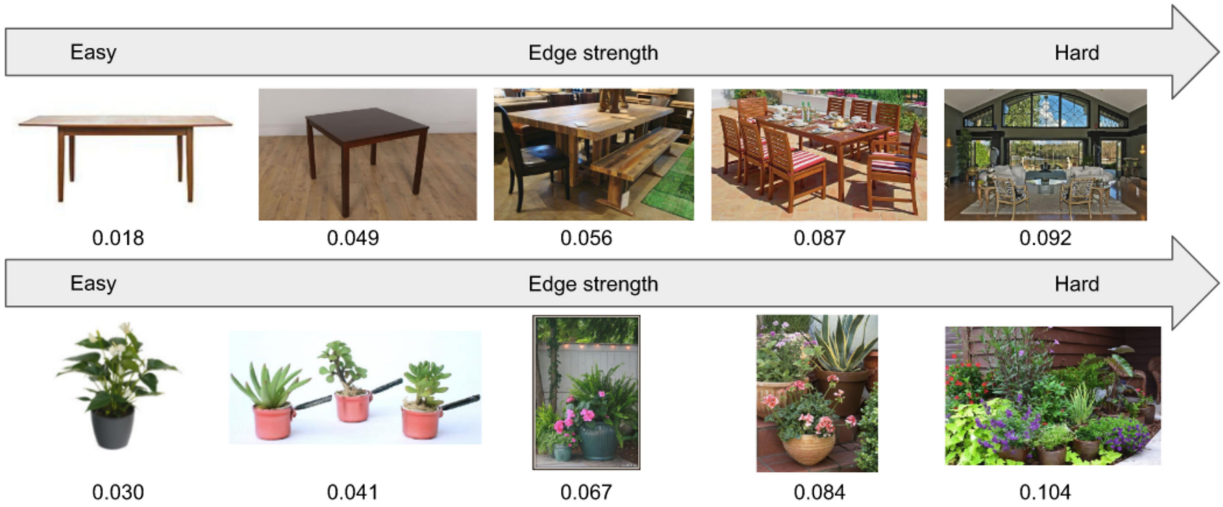

Fig. 5. Curriculum metric by the mean edge strength for web images (Top: table / Bottom: potted plant). The mean edge strength can reasonably represent the difficulty of images. Images with clean background and single object usually have small mean edge strength and images with complicated background and cluttered objects usually have large edge strength.

TABLE III
RESULTS OF THE DETECTION MAP ON VOC2007 TEST SET BY USING OUR CONSTRUCTED BING DATASET OR THE 'FLICKR CLEAN' DATASET AS EASY IMAGES AND VOC IMAGES AS HARD IMAGES FOR EASY-TO-HARD TRAINING

Web dataset	mAP
WebETH(Flickr clean)	35.5
WebETH(Bing)	**36.0**

our constructed web image dataset for WSD. As mentioned in Section III-B, we construct a web image dataset of 34 k images using Bing image search engine with attributes and related images. Considering that many selected web images are of high resolution, which causes huge complexity in the proposal generation process, we resize the longer side of all images to 600 pixels and keep the aspect ratios. We treat all web images as easy images and all VOC images as hard images. Simple web images are trained first followed by more complicated VOC images. In particular, we train the web images for 10 epochs. The learning rate is 10^{-5} for the first 5 epochs and 10^{-6} for the last 5 epochs. After that, VOC images are trained for 30 epochs: first 10 epochs with learning rate 10^{-5} and then decaying to 10^{-6}.

Table III shows the detection result of our method 'WebETH(Bing)' that exploits our constructed Bing dataset and trains the network in an easy-to-hard manner. Comparing Tables II and III, we can see that our method 'WebETH(Bing)' significantly improves the baseline 'WSDDN', increasing mAP from 33.9% to 36%, and also outperforms the VOC curriculum method 'CurrWSDDN'.

We also conduct experiments on another publicly available web dataset, STC Flickr clean dataset [30], which contains more than 40 k super clean images and has been proven to have good performance in generating good saliency maps to train weakly supervised segmentation networks. Surprisingly, including STC Flickr clean has no improvement over the VOC curriculum method 'CurrWSDDN', despite its result being much better than the baseline using only VOC images. In contrast, using our noisy Bing dataset 'WebETH(Bing)' beats both the VOC curriculum method 'CurrWSDDN' and the Flickr clean dataset 'WebETH(Flickr clean)'. This suggests that our approach of constructing a multi-attribute web dataset with large diversity is practically useful in this context.

We also explicitly study the difficulty distributions of Bing images, Flickr clean images and VOC2007 trainval images by performing statistical analysis using the image difficulty histograms (by percentage of images in the dataset). As shown in Fig. 6, although Flickr clean images are generally easier than VOC images, the difficulty distributions of Flickr dataset and VOC dataset are very similar. Whereas in Fig. 7, Bing dataset shows a large shift in the difficulty histogram. It shows that Bing dataset introduces a large number of extremely easy samples and these images can form a good complementary set to harder VOC images. Therefore, compared with 'CurrWSDDN' that uses curriculum learning scheme within VOC dataset, introducing Flickr clean data does not bring additional performance gain but involving Bing dataset can improve the results.

D. Results Regarding Relevance Metrics

Here we conduct experiments to study the effectiveness of using semantic relevance and distribution relevance. Fig. 4 gives some examples of the two relevance metrics. For the semantic relevance, we use the classification scores by the outlier detector described in Section III-B2, whose values vary from negative to more than 20. We set a semantic relevance threshold of 8 so that web images with scores lower than 8 are excluded. This prevents from mixing in noisy images without target objects into the early stage of training. For the distribution relevance, its relevance region includes web images which are members of top k-th nearest neighbors of one of VOC images, as illustrated in Fig. 4.

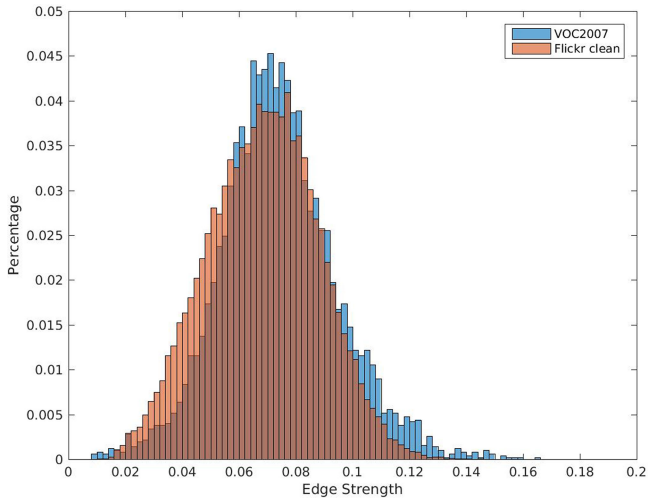

Fig. 6. Difficulty distributions of Flickr clean dataset [30] and VOC2007 trainval set: Percentage of images in the dataset at each difficulty range in Flickr dataset and in VOC dataset. The difficulty value is calculated by average edge strength of an image.

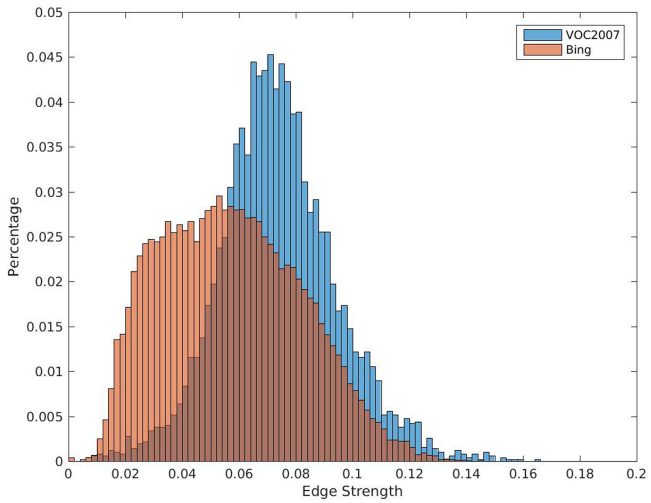

Fig. 7. Difficulty distributions of Bing dataset and VOC2007 trainval set.

TABLE IV
RESULTS OF THE DETECTION MAP ON VOC2007 TEST SET WITH DIFFERENT RELEVANCE METRICS AND USING OUR CONSTRUCTED BING DATASET WITH THE EASY-TO-HARD TRAINING

Transfer metrics	mAP
WebRelETH(Dist-Rel)	35.9
WebRelETH(Semantic-Rel)	**36.8**

Table IV shows the results using the two relevance metrics. We can see that with the semantic relevance, the detection result increases from 36.0% to 36.8%, whereas the kNN based distribution relevance (WebRelETH(Dist-Rel)) gives a slightly lower result, which suggests that similar images might not be always preferred. As a non-convex optimization problem, the training of WSD tends to drift to optimize small clusters of training samples. Although additional training instances with a similar distribution can help achieve lower training loss, it is not as helpful as involving new training samples with larger diversity, which leads to better generalization ability. This may also explain why STC Flickr clean dataset is not so helpful since the images in the Flickr clean dataset also have a similar distribution as VOC dataset.

E. More Comparison Results

We first illustrate the localization results along the training process with random initialization and the prior web model (see Fig. 8). Without the web initialization, the network is given a wrong initial detection of the object and fails to recover the correct object location in further training. With the web initialization, the top-scoring bounding box can gradually shift to a more accurate location along the training process.

Table V lists out the per-category average precision results of different WSD methods on VOC2007 test set with training on VOC2007 trainval set. It can be seen that compared with other existing WSD methods, the baseline method WSDDN achieves reasonably good performance. We would like to point out that our list in Table V might not be exhaustive since there might be some very recent WSD methods that report better performance. Since our solution is general, which can be added on top of any WSD baseline, it is more meaningful to evaluate our methods w.r.t the baseline.

The baseline WSDDN is applied on target dataset Pascal VOC2007 and web datasets including Flickr clean [30] and our Bing dataset. The results are listed in Table V. It is observed that using web data alone to train WSDDN is ineffective. This is due to the large domain mismatch between web data and target data. Web data lacks the samples with crowded or even occluded instances and also samples with complicated backgrounds. We conduct experiments on raw web dataset without relevance constraints and with semantic constraints to Bing images. Adding the relevance constraints leads to a 2.3% improvement. However, while purely training on web images, training with Flickr clean dataset alone leads to better test results than training with Bing dataset alone. This is because Flickr clean dataset is more similar to the target data and therefore it can adapt better in the test. However, by further training on target dataset that introduces more complicated images, the incremental gain is smaller for Flickr clean data but larger for Bing data in our proposed learning scheme.

Based on the original WSDDN framework, we consider five variants: using only VOC images with the curriculum regularizer (CurrWSDDN), simply combining our web images with VOC images with the semantic relevance for training (WebRel), combining our web images with VOC images for easy-to-hard training (WebETH), combining our web images with VOC images with the semantic relevance for easy-to-hard training (WebRelETH), and combining our web images with VOC images with the semantic relevance for easy-to-curriculum training (WebRelETC), where we train easy web images first and then train VOC images in a more detailed curriculum.

The results of CurrWSDDN, WebETH and WebRelETH have been discussed previously w.r.t. Tables II, III and IV, which demonstrate the effectiveness of the curriculum regularizer,

Fig. 8. Localization results with random initialization and web model initialization by taking a snapshot after every 4 epochs. We draw the top-scoring bounding box on the image to show the localization ability of the model at each training stage.

TABLE V
COMPARISONS OF THE DETECTION AVERAGE PRECISION RESULTS (%) ON VOC2007 TEST SET WITH TRAINING ON VOC2007 TRAINVAL SET. WE USE VGG-F MODEL PRETRAINED ON IMAGENET. WSDDN RESULTS ARE OBTAINED USING THE PUBLISHED CODE ON GITHUB WITH THE SAME SETTING STATED IN [4]. RESULTS OF OTHER METHODS ARE FROM THEIR PAPERS

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
Bilen et al. [2]	42.2	43.9	23.1	9.2	12.5	44.9	45.1	24.9	8.3	24.0	13.9	18.6	31.6	43.6	7.6	**20.9**	26.6	20.6	35.9	29.6	26.4
Bilen et al. [3]	46.2	46.9	24.1	16.4	12.2	42.2	47.1	35.2	7.8	28.3	12.7	21.5	30.1	42.4	7.8	20.0	26.8	20.8	35.8	29.6	27.7
Cinbis et al. [6]	39.3	43.0	28.8	20.4	8.0	45.5	47.9	22.1	8.4	33.5	23.6	29.2	38.5	47.9	20.3	20.0	35.8	30.8	41.0	20.1	30.2
Wang et al. [27]	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	30.9
Teh et al. [25]	48.8	45.9	37.4	**26.9**	9.2	50.7	43.4	43.6	10.6	35.9	27.0	38.6	48.5	43.8	**24.7**	12.1	29.0	23.2	48.8	41.9	34.5
ContextLocNet(contrastive S) [12]	**57.1**	52	31.5	7.6	11.5	55	53.1	34.1	1.7	33.1	**49.2**	42	47.3	56.6	15.3	12.8	24.8	**48.9**	44.4	47.8	36.3
WSDDN [4]	41.8	**57.7**	31.8	16.2	9.2	59.2	53.0	39.1	3.6	34.6	14.2	33.5	50.2	53.5	9.8	15.6	**37.3**	21.0	**53.1**	43.3	33.9
WSDDN(Bing only)	35.3	16.3	17.0	6.2	2.8	43.6	37.2	27.6	1.4	11.9	2.2	25.2	27.7	21.3	11.8	5.3	10.8	15.8	27.9	16.2	18.2
WSDDN(Bing rel only)	37.4	22.6	18.5	6.9	1.7	42.2	38.0	29.9	1.0	14.9	1.7	37.1	34.2	33.9	11.7	4.4	17.0	16.3	27.7	12.5	20.5
WSDDN(Flickr clean only)	31.4	26.6	22.0	10.0	1.5	43.0	38.1	36.6	1.7	12.3	19.7	32.8	34.1	38.6	8.4	5.7	17.6	29.5	32.0	18.2	23.0
CurrWSDDN	40.4	54.6	28.2	15.4	10.4	57.4	53.0	**44.5**	1.2	35.3	30.9	41.5	51.3	53.0	11.6	16.3	34.5	39.0	46.0	45.0	35.5
WebRel(ours)	40.7	51.5	31.0	10.7	10.0	61.0	43.2	39.4	1.8	30.1	35.5	**46.4**	52.3	50.6	9.0	13.4	30.4	31.8	41.2	42.3	33.6
WebETH(ours)	40.2	51.6	33.3	13.5	**13.0**	**62.8**	**54.5**	38.7	**11.8**	34.8	25.1	42.2	50.5	55.3	13.1	19.0	31.4	34.6	49.3	44.6	36.0
WebRelETH(Dist-Rel,ours)	38.4	48.2	33.5	11.4	11.5	62.2	53.4	32.9	8.7	37.3	38.8	34.6	50.5	54.8	9.6	16.9	33.8	45.2	48.2	48.4	35.9
WebRelETH(ours)	44.4	52.1	**38.1**	10.2	12.3	61.5	54.4	33.5	7.6	37.2	30.2	37.6	**55.4**	**57.3**	9.1	18.3	35.9	43.0	47.6	**50.0**	**36.8**
WebRelETC(ours)	38.9	52.4	33.4	11.2	10.5	59.9	53.8	36.4	3.0	**38.5**	41.8	38.8	53.9	56.0	11.9	18.9	35.1	43.2	46.2	47.2	36.6

TABLE VI
RESULTS OF THE DETECTION AVERAGE PRECISION (%) ON VOC2012 TEST SET WITH TRAINING ON VOC2012 TRAINING SET. WE USE VGG-F MODEL PRETRAINED ON IMAGENET. WSDDN RESULTS ARE OBTAINED USING THE PUBLISHED CODE ON GITHUB WITH THE SAME SETTING STATED IN [4]

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
WSDDN [4]	53.4	53.2	36.2	7.9	16.4	57.2	35.3	24.8	6.5	29.0	13.7	31.1	47.1	57.2	11.0	18.9	28.6	19.4	**42.3**	39.6	31.4
WebRelETH (ours)	**57.6**	55.1	**38.5**	**8.6**	20.4	**59.4**	**36.4**	33.6	**14.0**	**34.8**	21.7	**39.4**	**51.3**	**62.8**	11.5	**19.2**	**30.2**	**23.9**	41.2	**44.5**	**35.2**
WebRelETC (ours)	55.6	**56.2**	35.3	7.4	**20.5**	55.6	32.6	**34.8**	9.7	32.9	**32.1**	34.6	48.4	61.6	**15.5**	18.9	27.3	15.7	41.2	43.5	34.0

the constructed web dataset, the proposed relevance metrics, respectively. For WebRel, its result is even worse than the baseline WSDDN, which suggests that it is not an effective way to simply combine data from two sources. In our case, a large number of easy images dominate the training so that the model cannot be well trained for hard samples. For WebRelETC, we expect that the web images to have a similar difficulty level but VOC images need to be partitioned into more levels of difficulty. We first train on easy web images and adopt five-level curriculum regions for VOC images. It is found that its average precision performance is slightly worse than WebRelETH. This suggests that it is not always good to further break down the higher level curriculum for every class if the lower-level curriculum of simple web images have been used. Overall, our WebRelETH achieves the best mAP of 36.8%, outperforming the baseline by 2.9%.

Fig. 11 visualizes the detailed comparisons between WebRelETH and WSDDN. It can be seen that WebRelETH improves the detection results for most classes. Remarkably, it doubles the original WSDDN AP for "chair", "dining table" and "sofa" in VOC2007. Objects of these classes are considered hard to detect since they usually appear in a very complicated background with the coexistence of many other classes in the

Fig. 9. Visual results of WSDDN and our best model (WebRelETH). Our model can refine the bounding boxes as shown in the top two rows. Missing objects in the original model can also be detected in some test images as shown in the bottom rows.

TABLE VII
RESULTS OF THE DETECTION AVERAGE PRECISION (%) ON VOC2007 TEST SET. OICR [24] RESULTS ARE OBTAINED FROM THEIR PAPER. WE COMPARE WITH THEIR BEST MODEL (WITHOUT ENSEMBLE) USING VGG16

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
OICR-vgg16 [23]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	**15.7**	**24.1**	41.7	46.9	**64.3**	**62.6**	41.2
WebRelETH (ours)	**58.6**	**64.3**	**43.0**	**24.2**	**21.7**	**66.3**	**62.3**	**39.6**	**28.6**	**49.2**	**46.4**	**52.0**	**50.4**	**65.4**	10.0	**24.1**	**47.3**	**50.4**	58.3	58.1	**46.0**

target dataset. The target dataset lacks easy images to give a good prior knowledge of the appearances of these classes. Therefore, with the diversified easy web images introduced, we see a huge performance increase in these categories.

In Fig. 11, it is also observed that for WebRelETC, although the mean average precision is not as good as WebRelETH, it shows a huge improvement in "dining table" class and slight improvements in classes like "sofa" and "potted plant". These classes usually appear in the cluttered scene images and often with other object classes in the foreground and they are often considered hard classes in the object detection problem. For these classes, further breaking down the curriculum levels may help in learning the appearance model more steadily by gradually involving training images from the next difficulty level.

Table VI shows the experiment results for VOC2012. Our method also achieves up to 3.8% improvement in this dataset. Similar to VOC2007, WebRelETH outperforms WebRelETC, although WebRelETC excels largely in "dining table" by more than 10%. Fig. 9 gives some visual comparisons of the detection results using WSDDN and our best model (WebRelETH). It can be seen that our model can refine the bounding boxes (see the top two rows of Fig. 9), and missing objects in WSDDN can also be detected by our model in some test images (see the bottom rows of Fig. 9).

F. Advanced WSD Networks

To show that our method can be generalized to other WSD methods, we apply our learning scheme to the new state-of-the-art method OICR [24]. This method utilizes multiple online instance classifiers to refine the localization of the objects while training the WSD network. We incorporate our hierarchical curriculum learning scheme (WebRelETH) which uses the multi-attribute web dataset with semantic relevance constraints to train the network in [24]. Following the original OICR setting, the learning rate is set to 0.001 for the first 40000 iterations and then decreases to 0.0001. In total, we train 40000 iterations with web images and 70000 iterations with VOC 2007 trainval images. Each iteration contains two images. Our final results are compared with their best results (without ensemble) reported in their paper. In Table VII, we show that our approach achieves a significant improvement of 4.8% in mAP, suggesting that our learning scheme is general to different WSD networks and can achieve more significant performance boosts with a more advanced WSD network. Some visual results are shown in Fig. 10.

In addition to the generalization ability for different WSD networks, we also show that the web prior knowledge is agnostic to the target datasets. We test our model trained on VOC 2007 using VOC 2012 validation set and achieve 5.2% improvement in mAP compared with the baseline model that does

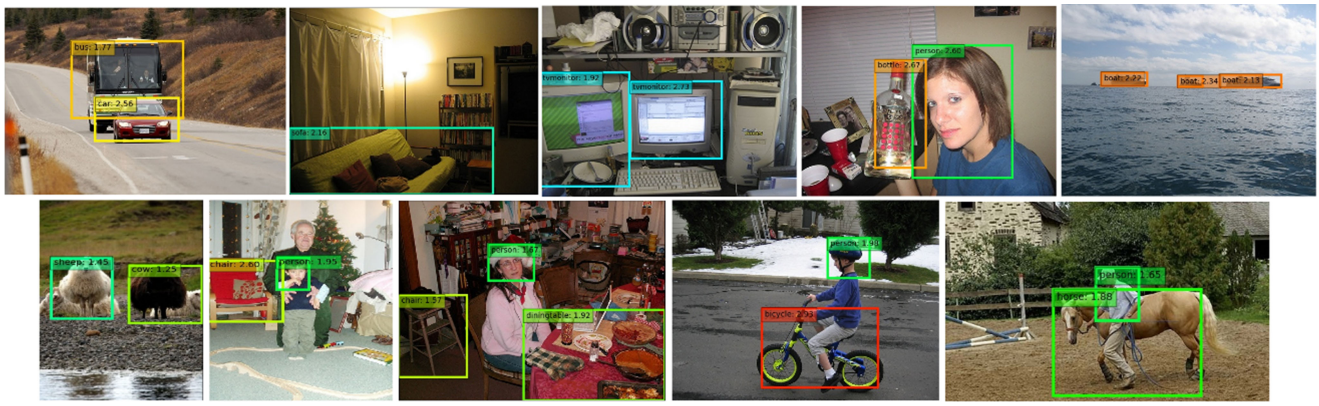

Fig. 10. Visual results of WebRelETH with OICR [24]. The first row shows some good examples. The second row shows some failure cases, where the detector fails to distinguish very similar categories such as "sheep" and "cow". The performance has a notable drop in "person" category where the detected regions mainly focus on the face of person. This could be because that the web data have a large percentage of profile images with only human face instead of the whole body. Moreover, deeper networks such as vgg16 tend to focus on learning objects parts.

TABLE VIII
RESULTS OF THE DETECTION MAP ON VOC2012 VAL SET WITH MODELS TRAINED WITH VOC2007 TRAINVAL SET

Methods	mAP
OICR-vgg16	37.6
WebRelETH (ours)	**42.8**

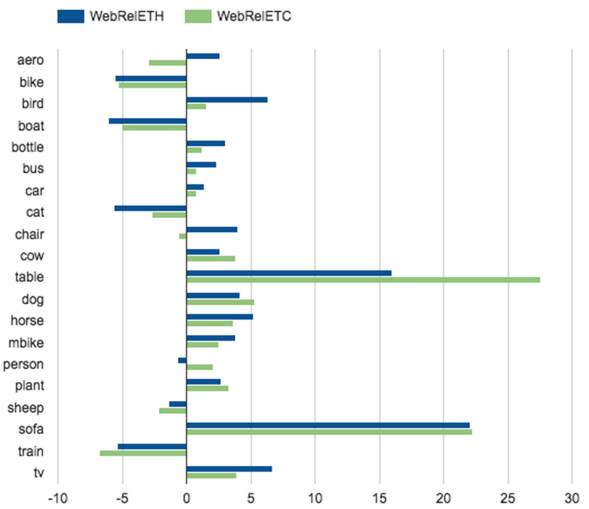

Fig. 11. Increment in detection AP (%) using our best models in comparison with original WSDDN on VOC2007.

not explore web data in training. This further shows the advantage of exploiting general and diverse web data as prior information.

V. CONCLUSION

This paper has addressed two questions: how to construct a large, diverse and relevant web image dataset and how to use it to help weakly supervised object detection. Particularly, for constructing the web dataset, we introduced a sophisticated expand-to-condense process to first expand web data with attributes and related images and then condense the dataset with semantic relevance or distribution relevance. For helping the target dataset, we applied an easy-to-hard learning scheme. Extensive results have validated that our easy-to-hard learning with web data is effective and the multi-attribute web data do help in training a weakly supervised detector.

REFERENCES

[1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, ACM, 2009, pp. 41–48.
[2] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with posterior regularization," in *Proc. Brit. Mach. Vis. Conf.*, 2014, vol. 3.
[3] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1081–1089.
[4] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2846–2854.
[5] X. Chen and A. Gupta, "Webly supervised learning of convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1431–1439.
[6] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 39, no. 1, pp. 189–203, 2017.
[7] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 914–922.
[8] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1, pp. 31–71, 1997.
[9] S. K. Divvala, A. Farhadi, and C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3270–3277.
[10] J. Fu *et al.*, "Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1985–1993.
[11] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, vol. 2, pp. 2694–2700.
[12] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 350–365.
[13] J. Krause *et al.*, "The unreasonable effectiveness of noisy data for fine-grained recognition," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 301–320.
[14] K. Kumar Singh, F. Xiao, and Y. J. Lee, "Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3548–3556.
[15] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Vis. Pattern Recognit., Comput. Soc. Conf.*, 2006, vol. 2, pp. 2169–2178.

[16] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3512–3520.
[17] S. Li, X. Zhu, Q. Huang, H. Xu, and C.-C. J. Kuo, "Multiple instance curriculum learning for weakly supervised object detection," in *Proc. Brit. Mach. Vis. Conf.*, 2017.
[18] Y. Li *et al.*, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.
[19] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comp. Vis.*, Springer, 2016, pp. 21–37.
[20] L. Niu, W. Li, D. Xu, and J. Cai, "Visual recognition by learning from web data via weakly supervised domain generalization," *IEEE Trans. Neural Netw. Learn.*, vol. 28, no. 9, pp. 1985–1999, Sep. 2017.
[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[22] S. Reed *et al.*, "Training deep neural networks on noisy labels with bootstrapping," 2014, arXiv:1412.6596.
[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[24] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2843–2851.
[25] Y. Tang, X. Wang, E. Dellandréa, and L. Chen, "Weakly supervised learning of deformable part-based models for object detection via region proposals," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 393–407, Feb. 2017.
[26] Q. Tao, H. Yang, and J. Cai, "Zero-annotation object detection with web knowledge transfer," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 387–403.
[27] E. W. Teh, M. Rochan, and Y. Wang, "Attention networks for weakly supervised object localization," in *Proc. Brit. Mach. Vis. Conf.*, 2016.
[28] R. T. Ionescu *et al.*, "How hard can it be? estimating the difficulty of visual search in an image," in *Proc. IEEE Conf. Comput. Visi. Pattern Recognit.*, 2016, pp. 2157–2166.
[29] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 431–445.
[30] Y. Wei *et al.*, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2016.
[31] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2691–2699.
[32] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Augmenting strong supervision using web data for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2524–2532.
[33] Y. Yao *et al.*, "Exploiting web images for dataset construction: A domain robust approach," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1771–1784, Aug. 2017.
[34] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 391–405.

Qingyi Tao received the B.S. degree from the School of Computer Science and Engineering, Nanyang Technological University, Singapore. She is currently working toward the Ph.D. degree with NTU. She is currently a Deep Learning Solutions Architect with NVIDIA AI Technology Center. Her research interests include computer vision and deep learning.

Hao Yang received the B.Sc. degree from Shanghai Jiao Tong University, Shanghai, China, in 2011, and the Ph.D. degree from the School of Computer Science and Engineering, Nanyang Technological University, Singapore, in 2016 under the supervision of Prof. Jianxin Wu and Prof. Jianfei Cai. He is an Applied Scientist with Amazon, Seattle, WA, USA. His research interests include computer vision and machine learning, especially large-scale visual recognition systems.

Jianfei Cai (S'98–M'02–SM'07) received the Ph.D. degree from the University of Missouri-Columbia, Columbia, MO, USA. He is currently a Professor and has served as the Head of Visual and Interactive Computing Division and the Head of Computer Communication Division, School of Computer Science and Engineering, Nanyang Technological University, Singapore. He has authored or coauthored more than 200 technical papers in international journals and conferences. His major research interests include computer vision, multimedia, and deep learning. He is currently an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, and has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, and TRANSACTIONS ON CIRCUIT AND SYSTEMS FOR VIDEO TECHNOLOGY.