Learning Object Detectors With Semi-Annotated Weak Labels

Dingwen Zhang, Junwei Han, *Senior Member, IEEE*, Guangyu Guo, and Long Zhao

***Abstract*— For alleviating the human labor associated with annotating the training data for learning object detectors, recent research has focused on semi-supervised object detection (SSOD) and weakly supervised object detection (WSOD) approaches. In SSOD, instead of annotating all the instances in the whole training set, people only need to annotate the part of the training instances using bounding boxes. In WSOD, people need to annotate the image-level tags on all training images to indicate the object categories contained by the corresponding images since more detailed bounding box annotations are no longer needed. Along this line of research, this paper makes a further step to alleviate the human labor in annotating training data, leading to the problem of object detection with semi-annotated weak labels (ODSAWLs). Instead of labeling image-level tags on all training images, ODSAWL only needs the image-level tags for a small portion of the training images, and then, the object detectors can be learned from a small portion of the weakly-labeled training images and from the remaining unlabeled training images. To address such a challenging problem, this paper proposes a cross model co-training framework that collaborates an object localizer and a tag generator in an alternative optimization procedure. Specifically, during the learning procedure, these two (deep) models can transfer the needed knowledge (including labels and visual patterns) between each other. The whole learning procedure is accomplished in a few stages under the guidance of a progressive learning curriculum. To demonstrate the effectiveness of the proposed approach, we implement the comprehensive experiments on three benchmark datasets, where the obtained experimental results are quite encouraging. Notably, by using only about 15% weakly labeled training images, the proposed approach can effectively approach, or even outperform, the state-of-the-art WSOD methods.**

***Index Terms*— Computer vision, image processing, object detection, learning (artificial intelligence).**

I. Introduction

AS A FUNDAMENTAL problem in the computer vision community, object detection [1], [2] has received wide attention in the last few decades. The goal is to find objects of interest in images via assigning labels to the extracted bounding box regions. In light of the rapid development of the deep learning techniques, especially the convolution neural networks (CNNS) (e.g., [3], [4]), the modern object detection approaches (e.g., [5]–[8]) can already perform promising object detection in real world images or videos. Most of such object detection approaches require large amounts of instance-level labeled training data (i.e., the bounding box annotation and the corresponding semantic tags) and thus they are treated as the fully supervised object detection approaches (see the first line of Table I). However, the manual annotation of such training data is highly time-consuming and labor intensive. To this end, approaches to alleviate the human labor of annotating the training data for learning object detectors have received great research interests in recent years.

TABLE I

Comparison of the Manual Annotations Needed for Different Schools of Object Detection Approaches

	(Image) tag	Bounding box
Fully supervised object detection	All	All
Semi-supervised object detection	Part	Part
Weakly supervised object detection	All	None
Object detection with semi-annotated weak labels	Part	None

Among the existing works, one school of the approaches to alleviate the human labor in training object detectors is the semi-supervised object detection (SSOD) methods (such as [9] and [10]) where people only need to annotate a part of the training instances using bounding boxes instead of annotating all the instances in the whole training set (see the second line of Table I). Another school of approaches is the weakly supervised object detection (WSOD) methods (such as [11] and [12]) where people need to annotate the image-level tags on all training images to indicate the object categories contained by the corresponding images, while more detailed bounding box annotations are not needed any more (see the third line of Table I).

Along this line of research, this paper makes a further step to alleviate the human labor in annotating training data. As shown in the bottom line of Table I, instead of labeling image-level tags on all training images, we further weaken the human labor by only requiring people to provide the image-level weak annotations on a small portion of the training images, leading to the investigated problem of object detection with semi-annotated weak labels (ODSAWL). In ODSAWL, the discriminative patterns of different object semantics need

Manuscript received May 1, 2018; revised September 15, 2018; accepted November 25, 2018. Date of publication November 30, 2018; date of current version December 6, 2019. This work was supported in part by the National Key R&D Program of China under Grant 2017YFB0502904, in part by the National Science Foundation of China under Grants 61876140 and 61773301, and in part by the China Postdoctoral Support Scheme for Innovative Talents under Grant BX20180236. This paper was recommended by Associate Editor G. Hua. *(Corresponding author: Junwei Han.)*

D. Zhang is with the School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: zhangdingwen2006yyy@gmail.com).

J. Han, G. Guo, and L. Zhao are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

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

Digital Object Identifier 10.1109/TCSVT.2018.2884173

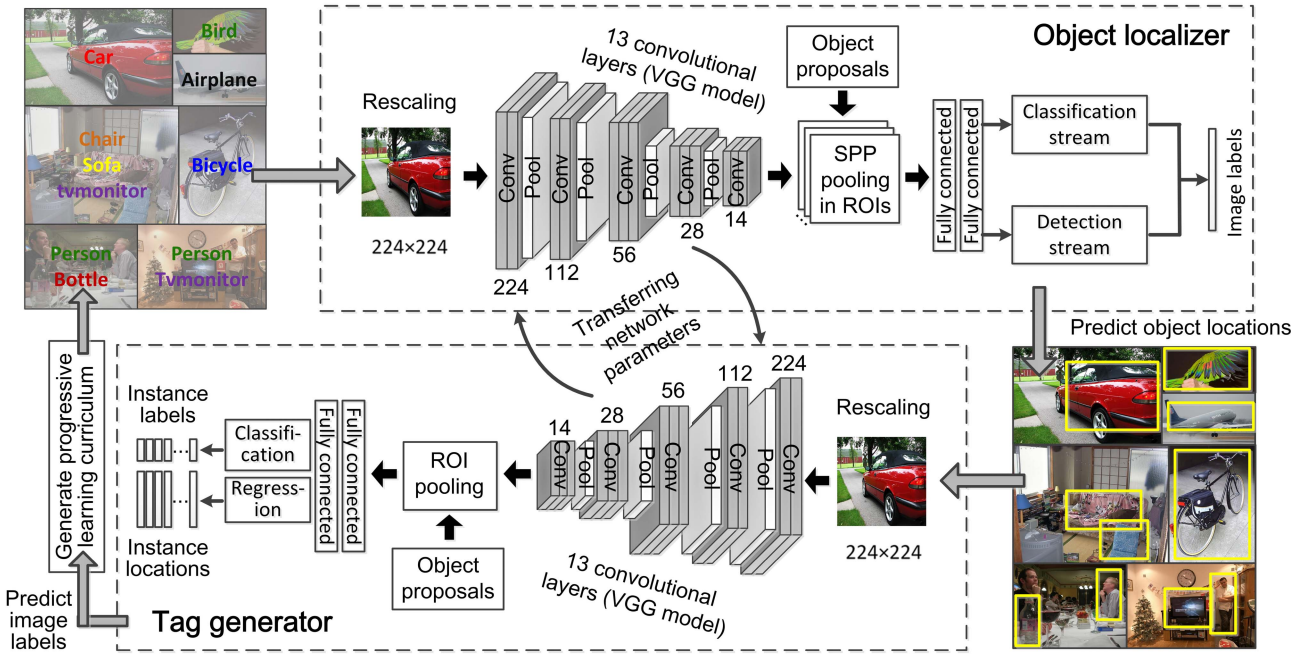


Fig. 1. The proposed framework for learning object detectors with semi-annotated weak labels. It contains a object localizer to transfer knowledge from image tag to object location (see Sec. III-B1), a tag generator to transfer knowledge from object location to image tag (see Sec. III-B2), and a progressive learning curriculum to guide a steady and robust learning procedure under the extremely weak supervision (see Sec. III-B3).

to be captured and the corresponding object detectors need to be learnt from the small portion of the weakly-labeled training images and the rest unlabeled training images. Thus, ODSAWL tends to be more challenging than the existing SSOD and WSOD approaches as the learning ambiguity we are facing in ODSAWL is much larger than what we met before. Albeit challenging, ODSAWL is of great research significance as it can not only further alleviate the human labor in learning object detectors, but also provide a potential way to leverage the 'unlimited' and 'free' unlabeled training data.[1]

To address the challenging ODSAWL problem, we propose a novel cross model co-training framework, where an object localizer and a tag generator are trained alternatively and iteratively to learn the faithful discriminative visual patterns of different object categories. Specifically, provided by the image-level tags, the object localizer learns to predict the category-specific instance-level labels for each bounding box region, thus localizing objects of interest with bounding box annotation. With such instance-level labels, the tag generator first learns to pursue the stronger classification on the bounding box instances and then uses such bounding box classification predictions to generate reliable image tags for the unlabeled training images. The generated image tags can in turn be used to refine the object localizer. In this way, the object localizer and tag generator can work collaboratively by transferring the needed knowledge (the learnt visual patterns and the needed training labels) to each other. To guide a steady and robust cross model co-training procedure, we build a progressive learning curriculum according to the generated image-level tags on the unlabeled training images and use it to select relatively confident training data among each learning stage. With such cross model co-training framework, the proposed approach is able to collaborate two function distinct learning models to work compatibly and robustly to overcome the serious learning ambiguity issue inherent in learning with the semi-annotated weak labels.

The concrete framework built (by us) for learning object detectors with semi-annotated weak labels is shown in Fig. 1, where the network architecture of the object localizer is shown in the top-right block and the network architecture of the tag generator is shown in the bottom-left block. Due to the similar network architectures shared by these two learning models, the visual patterns learnt by each individual one can be easily transferred to the other through transferring a part of the network parameters. Doing so can reduce the chance to learn the contradictory patterns from different models during the learning procedure, and thus better guide the interaction and collocation during the cross-model co-training procedure. After each learning stage, we generate a learning curriculum by estimating the confidence of the obtained image tags on the unlabeled training images. According to the built learning curriculum, only the part of unlabeled training images with relatively confident image tag predictions will be used in the next learning stage. The whole learning procedure performs such learning stages round by round until reaching the termination condition. During the learning procedure, the gradually involved training data forms a progressive learning curriculum to guide the learner to get rid of the learning ambiguity hoarded in unconfident training images, thus leading to a steady and robust learning procedure.

There are two previous works [13], [14] that are closely related to this paper. In [13], an Expectation-Maximization

[1]Here the 'unlimited' and 'free' training data refers to visual data that grows fast and can be easily accessed, e.g., the online images or videos.

Algorithm-based weakly- and semi-supervised object detection framework is established. The goal is to leverage both the weak image tags on all training images and a small number of strongly labeled bounding box annotations for training the corresponding object detectors. However, this work is not able to solve our problem as there is no acquirable strong bounding box annotations in our problem. Another work close to this proposed work is proposed by Zhang *et al.* [14], where a self-paced curriculum learning-based weakly supervised object detection framework is established. Its goal is consistent with the conventional WSOD approaches, i.e., learning object detectors using all the weakly labeled training images. Zhang *et al.* [14] build a novel framework to start training the object detectors from a small number of images and samples, which is considered as the confident ones by their method, and then enhance the object detectors by involving all training samples in all training images. However, the intrinsic difference between the problems in our paper and [14] is that only a small part of the training images (instead of all of them) are weakly labeled in our scenario. In addition, we cannot adopt the way to select a part of data from the weakly labeled training images like [14] as such training data is insufficient for training powerful deep object detectors.

To sum up, there are mainly three-fold contributions in this paper:

1) This paper investigates a pioneer research direction towards learning object detectors with extremely weak supervision. It is of great research significance as it can not only further alleviate the human labor in annotating training data for learning object detectors, but also provide a potential way to leverage the 'unlimited' and 'free' unlabeled training data.

2) This paper designs a brand new cross model co-training framework, where two deep models, i.e., the object localizer and the tag generator, are collaborated in an alternative optimization procedure. In our framework, these two models can generate the training labels needed for each other and transfer the learnt visual patterns between them. Under the guidance of the established progressive learning curriculum, the proposed learning framework can work steady and robustly in the investigated ODSAWL problem.

3) Comprehensive experiments on three benchmark datasets are conducted to demonstrate the effectiveness of the proposed approach. Notably, by only using about 15% weakly labeled training images, the proposed approach can approach to, or even outperform, the state-of-the-art WSOD methods.

The rest of this paper is organized as follows. Section II reviews previous works related to this paper. Section III introduces the detailed techniques of the proposed approach. Section IV reports the experimental results for both evaluating the considered components in the designed framework and comparing the proposed approach with the state-of-the-art WSOD approaches. Finally, conclusions are drawn in Section V.

II. Related Works

A. Semi-Supervised Object Detection

In semi-supervised object detection (SSOD), e.g., [9], [10], [15], [16], only a part of the training instances are required to be labeled manually (with the corresponding bounding boxes and semantic tags) while the learner needs to leverage such limited annotation to infer the instance labels on the unlabeled training images and learn final object detectors based on both the labeled training images and unlabeled training images. Specifically, in order to constrain the semi-supervised learning process to avoid semantic drift, Misra *et al.* [9] proposes to combine multiple weak cues (appearance, motion, temporal etc.) in videos and exploit the decorrelated errors by modeling data in multiple feature spaces. Tang *et al.* [16] leverages knowledge from both visual and semantic domains to adapt an image classifier to an object detector in a semi-supervised learning setting. They demonstrate the combination of knowledge from different domains is important for improving SSOD.

Although SSOD can linearly reduce the human labor of annotating training data when compared with the conventional fully supervised object detection scheme, it still needs human to predivide bounding-box level annotations for the training images. On the contrary, the learning problem investigated in this paper only needs human to provide image-level tags on a small part of training images, which can alleviate more human labor as compared with SSOD.

B. Weakly-Supervised Object Detection

Different from SSOD, weakly-supervised object detection (WSOD), e.g., [11], [12], [17]–[22], needs people to annotate the image-level tags on all training images to indicate the object categories contained by the corresponding images, while more detailed bounding box annotations are not needed any more. Among these methods, Bilen and Vedaldi [20] proposes the first end-to-end weakly supervised object detection framework (named as WSDDN), which modifies the architectur of the conventional fully supervised object detection network to operate at the level of image regions and performs simultaneously region selection and classification. Jie *et al.* [22] proposes a deep self-taught learning approach, where the seed sample acquisition method is adopted to improve the location quality of the inferred positive instances as well as the learning quality of the subsequent network retraining process.

WSOD could alleviate more human labor than SSOD as no bounding-box annotation is needed. Whereas, compared with WSOD, the investigated ODSAWL framework can further linearly reduce the human labor of annotating training data as only the weak labels on a part of training images are needed to be provided.

C. Co-Training

The conventional co-training frameworks, such as [23]–[25], are usually used in semi-supervised learning scenario. These approaches have three assumptions: 1) features of the training data can be split into two sets; 2) each subfeature set is sufficient to train a good classifier; and 3) the two sets are conditionally independent given the class. However, the cross model co-training framework proposed in this paper is not same as these conventional co-training frameworks. Firstly, the problem investigated in this paper

(which can be seen as labeling partial training data with weak labels) is much more challenging than semi-supervised learning problem. More importantly, given the complex and challenging nature of the investigated ODSAWL problem, none of the aforementioned assumptions can be satisfied in our case. Thus, the conventional co-training frameworks cannot be used to solve our problem. It is also worth mentioning that, albeit different, the proposed co-training framework shares a consistent idea with the conventional co-training frameworks, i.e., the two learning models need to agree on as much training data as possible to ensure they can work compatibly and provide informative knowledge to each other. This is the reason why we need to transfer network parameters of these two models during the learning procedure.

III. The Proposed Approach

A. The Overall Learning Framework

In the investigated ODSAWL problem, we are only given a small number of weakly labeled training images $\{\mathcal{I}^l, \mathcal{T}^l\}$ ($\mathcal{I}^l$ indicates the images and $\mathcal{T}^l$ indicates the weak labels, i.e., image tags) and a number of unlabeled training images $\mathcal{I}^u$, while the final goal is to learn object detectors $\mathbf{W}$ to detect the objects of interest in the test image set $\mathcal{X}^t$. Let us denote the object proposals, i.e., instances, extracted in the training images and test images as $\mathcal{X}^{tr}$ and $\mathcal{X}^{te}$, respectively. The desired object detectors need to have discriminative capacity in separating the corresponding object category of a certain instance from other object categories as well as the image background.

To solve this problem, we cannot directly follow the frameworks proposed by the previous WSOD approaches as not all training images are labeled with image-level tags and training only using the part of weakly labeled training images won't be sufficient for obtaining strong enough object detectors. In addition, we cannot follow the frameworks proposed by the previous SSOD approaches as well, because the training data are partially labeled with the weak image-level tags rather than the strong instance-level bounding box annotations. Under this circumstance, we propose to collaborate two learning models with deep network architectures, i.e., the object localizer and the tag generator, in a novel co-training framework, where the object localizer learns discriminative visual patterns from the weakly labeled training images, while the tag generator further enhances the discrinimality of the learnt visual patterns to obtain stronger instance-level object detectors.

During the training procedure, we first train the object localizer under the supervision of $\{\mathcal{I}^l, \mathcal{T}^l\}$, and then use the obtained model to classify the instances in each training image, thus generating the instance-level labels for the instances in the training images, i.e., $\mathcal{Y}^{tr}$. Afterwards, we train the tag generator under the supervision of $\{\widetilde{\mathcal{X}}^{tr}, \widetilde{\mathcal{Y}}^{tr}\}$, which is the subset selected from $\{\mathcal{X}^{tr}, \mathcal{Y}^{tr}\}$. Next, we use the obtained model to predict the labels of each instance and further generate the image tags to the unlabeled training images. By estimating the confidence of the obtained image tags for each unlabeled training image, we selectively merge the part of training images with the relatively confident image tags to form the new weakly labeled training set $\{\mathcal{I}^l, \mathcal{T}^l\}$, which will be used in the next learning stage. The above process is iterated until reaching the termination condition (typically 3 to 4 stages suffice). The whole learning procedure is shown in Fig. 1 and Algorithm 1.

Algorithm 1: The Cross Model Co-Training Algorithm to Learn Object Detectors With Semi-Annotated Weak Labels

input : The weakly labelled training image set $\{\mathcal{I}^l, \mathcal{T}^l\}$ and unlabelled training image set $\mathcal{I}^u$;
output: The learnt object detectors $\mathbf{W}$;

1 Extract object proposals to form the instance set $\mathcal{X}^{tr}$;
2 **for** *t in* $1 \cdots T$ **do**
3 Transfer the body net parameters from the tag generator (if $t > 1$); Train the object localizer Φ under the supervision of $\{\mathcal{I}^l, \mathcal{T}^l\}$; Use the obtained Φ to generate the instance-level label set $\mathcal{Y}^{tr}$; Select the subset of training instances from $\{\mathcal{X}^{tr}, \mathcal{Y}^{tr}\}$ to form $\{\widetilde{\mathcal{X}}^{tr}, \widetilde{\mathcal{Y}}^{tr}\}$;
4 Transfer the body net parameters from the object localizer; Train the tag generator $\mathbf{W}$ under the supervision of $\{\widetilde{\mathcal{X}}^{tr}, \widetilde{\mathcal{Y}}^{tr}\}$; Use the obtained $\mathbf{W}$ to predict the labels of each instance; Generate the image tags to the unlabelled training images;
5 Estimate the learning confidence of unlabelled training images according to the obtained image tags; Merge the training images with the relatively confident image tags to form the new weakly labelled training set $\{\mathcal{I}^l, \mathcal{T}^l\}$;
6 **end**
7 **return** The learnt object detectors $\mathbf{W}$.

B. Cross Model Co-Training

In this subsection, we will introduce the proposed cross model co-training process in detail. As shown in Algorithm 1, there are mainly three steps in each learning stage of the proposed co-training framework, which are the knowledge transfer from image tag to object location (i.e., Step 3 in Algorithm 1), knowledge transfer from object location to image tag (i.e., Step 4 in Algorithm 1), and progressive learning curriculum generation (i.e., Step 5 in Algorithm 1), respectively. Next, we will describe each of these main steps.

1) Knowledge Transfer From Image Tag to Object Location: Given the small number of weakly labeled training image set $\{\mathcal{I}^l, \mathcal{T}^l\}$, we transfer the useful knowledge from the provided image tags to the object locations by adopting an object localizer. A brief network architecture of the object localizer is shown in Fig. 2 (a), which is designed based on the WSDDN model [20]. Essentially, the entire network architecture of the object localizer can be separated into three parts, i.e., the body net, neck net, and head net, respectively. The body net of the object localizer is used to extract informative visual patterns for the whole network, the neck net is used to extract spatial pyramids features of each object proposal region from the feature maps extracted by the body net, and the head net

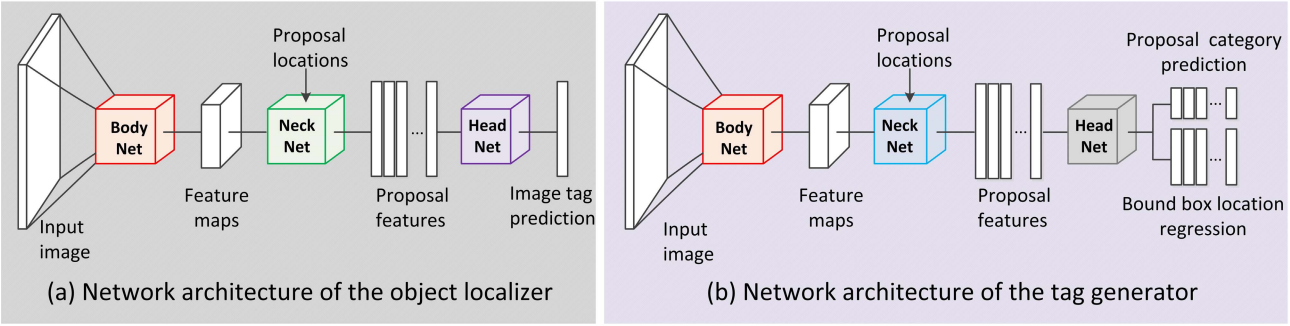


Fig. 2. Brief illustration of the network architectures of the object localizer (a) and tag generator (b). Notice that these two deep models share the same body net architecture while with different network architectures of their neck nets and head nets. More detailed description on these network architectures can be referred to in the main text.

is used to generate the final image tag predictions based on the obtained object proposal features. The concrete network architecture of the object localizer can be referred to in Fig. 1 and Fig. 2 (a).

Suppose there are N_l images in the weakly labeled training set, i.e., $\mathcal{I}^l = \{I_1, I_2, \cdots, I_{N_l}\}$, $\mathcal{T}^l = \{T_1, T_2, \cdots, T_{N_l}\}$. Denoting the network parameters in our object localizer as Φ. As in [20], the learning objective function contains a learning loss term and a learning regularizer term. The learning loss term is used to penalize predictions inconsistent with the given image labels. The learning regularizer is used to penalise the feature map discrepancies at the proposal features between the highest scoring proposal region and its nearby proposal regions. The training process uses stochastic gradient descent with momentum to optimise such objective function.

After the learning process, we use the obtained network parameters Φ to generate the prediction results, i.e., the C-dimensional vectors (C indicates the number of object categories that need to be learnt), of each object proposal region. Specifically, we feed forward the network to the penultimate layer, i.e., the element-wise production layer in the head net, and thus obtain the instance-level label set $\mathcal{Y}^{tr}$. Then, for training the tag generator with reliable supervision, we select a subset of confident proposals $\{\widetilde{\mathcal{X}}^{tr}, \widetilde{\mathcal{Y}}^{tr}\}$ from $\{\mathcal{X}^{tr}, \mathcal{Y}^{tr}\}$. Specifically, for each weakly labeled training images, we select one instance for each object category contained in the labeled image tag, which has the highest prediction score for the corresponding object category. Thus, the number of the training instances selected from each weakly labeled training image equals to the number of object categories contained in the corresponding image tag. For each unlabeled training image, we only select one instance that has the highest prediction score within it. The label of this instance is the object category with the highest prediction score.

2) Knowledge Transfer From Object Location to Image Tag: Given the confident proposal set $\{\widetilde{\mathcal{X}}^{tr}, \widetilde{\mathcal{Y}}^{tr}\}$ obtained based on the object localizer, we then transfer the useful semantics from such labeled proposals (labeled with the pseudo object categories accordingly) to generate image tags for the unlabeled training images by training a tag generator that built by the Fast R-CNN model [6]. A brief network architecture of the tag generator is shown in Fig. 2 (b).

Similar to the network architecture of the object localizer, the entire network architecture of the tag generator can also be separated into the parts of body net, neck net, and head net. The body net of the tag generator is designed the same as it in the object detector, which is also used to extract informative visual patterns for the whole network. The neck net is used to extract the features of each object proposal region from the feature maps extracted by the body net. In the head net of the tag generator, each of the proposal features extracted by the neck net is passed through two stream prediction layers and finally obtains its prediction scores of each object category and the bounding box location offset that can be used to better modify its original bounding box location. The concrete network architecture of the tag generator can be referred to in Fig. 1 and Fig. 2 (b).

Suppose there are N proposal instances extracted in the training image set $\{X^{tr}, Y^{tr}\}$, i.e. $\mathcal{X}^{tr} = \{x_1, x_2, \cdots, x_N\}$, $\mathcal{Y}^{tr} = \{y_1, y_2, \cdots, y_N\}$. Here $y_j, j \in [1, N]$ is redefined as the $C+1$-dimensional (C object categories plus one background category[2]) label vector of each proposal instance, which can be obtained by comparing its bounding box location with the bounding box locations of the proposals in $\widetilde{\mathcal{X}}^{tr}$ based on the IOU overlap criteria. If the bounding box locations have more than 50% overlap, y_j would be assigned with the object category of the corresponding proposal (have the largest overlap with it) in $\widetilde{\mathcal{X}}^{tr}$.

According to the obtained $\mathcal{Y}^{tr}$, each training proposal is associated with a pseudo ground-truth class. We also have the bounding-box regression target by comparing the proposal location with its corresponding pseudo ground-truth location. Then, we follow [6] to train the parameters $\mathbf{W}$ in our tag generator by minimizing a multi-task loss function, including a classification loss to recognize image content and a regression loss to learn bounding-box offsets.

After the learning process, we use the obtained network parameters $\mathbf{W}$ to generate the prediction results, i.e., the $C+1$-dimensional vectors $\mathbf{p}$, for each object proposal region. Notice that under this circumstance, the bounding box locations of each object proposal have been modified by the network. For a certain unlabeled training images $I_i^u \in \mathcal{I}^u$, we obtain its image tag $T_i^u = (T_{i,1}^u, T_{i,2}^u, \cdots, T_{i,C}^u)$ by:

$$T_{i,c}^u = \max(\max_{\mathbf{p} \in I_i^u} p_c - \operatorname*{mean}_{\mathbf{p} \in I_i^u} p_0, 0), \tag{1}$$

[2]The index of the background category is 0.

Fig. 3. Some examples of the generated image tags for the unlabeled training images. The left part shows the images labeled with accurate image tags, while the right part shows the images labeled with inaccurate image tags. The predicted image tags are shown in black while the true image tags are shown in red.

where $\mathbf{p} \in I_i^u$ indicates the prediction vectors of the proposal regions extracted in I_i^u. From (1), we can see that, in the obtained image tag, the object category prediction scores that are even smaller than the average background score would be considered as 0.

3) Generate Progressive Learning Curriculum: After obtaining the image tags of the unlabeled training images, we use them to generate the training data for learning the stronger object localizer in the next learning stage. Due to the fact that the obtained image tags cannot perfectly label the image content of all the unlabeled training images, directly using all of them would unavoidably involve much imprecise or even totally wrong supervision to the learning process of the object localizer and thus obtain trivial solutions in the end. To address this problem, we propose to establish a progressive learning curriculum to guide the learning procedure. Based on our observation (see Fig. 3), the image tags generated by the tag generator would have satisfactory accuracy in some of the unlabeled training images (as shown in the left part of Fig. 3) while unsatisfactory accuracy in some other unlabeled training images (as shown in the right part of Fig. 3). This inspires us to find ways to estimate the reliability of each unlabeled training image based on the corresponding image tag, and then implement the subsequent learning process of the object localizer on the reliable training images.

To achieve this goal, we consider two factors to estimate the learning reliability of each unlabeled training image. The first one is the image complexity. Intuitively, we can count the number of the non-zero elements in the obtained image tag:

$$IM_i^u = ||T_i^u||_0, \tag{2}$$

to estimate the complexity of the content of the i-th unlabeled training image, because IM_i^u can reflect the number of object categories that appear in the corresponding image and the images containing many different object categories are usually have complex image content (see Fig. 4).

The second factor is the tag confidence TC_i^u, which can be calculated by

$$TC_i^u = \frac{||T_i^u||_1}{||T_i^u||_0}, \tag{3}$$

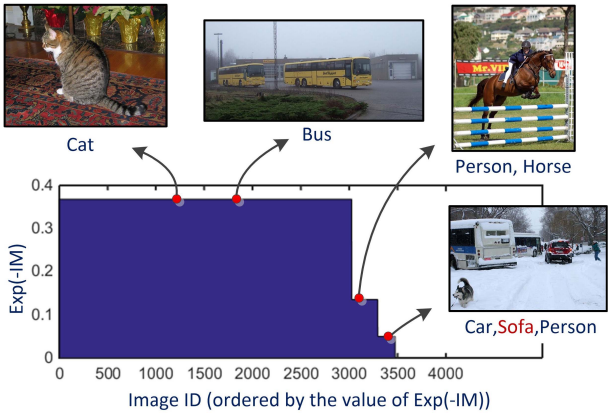


Fig. 4. Examples to illustrate the rationality of building the learning reliability based on the image complexity (IM). Object categories under each example are predicted by the learner, where blue ones are correctly predicted while the red one is wrongly predicted. From this figure, we can observe that examples with higher $\exp(-IM_i^u)$ values tend to contain relatively simpler image scenes and thus lower image complexity, while examples with lower $\exp(-IM_i^u)$ values tend to contain more complex image scenes and thus higher image complexity. Besides, in images containing more object categories, it is harder to precisely predict all the object categories (see the example on the far right).

which is essentially the average prediction value of the object categories contained by the i-th unlabeled training image. It is easy to observe that the images with larger TC_i^u would be assigned with more confident image tag (see Fig. 5).

By considering the aforementioned two factors, we estimate the learning reliability of each unlabeled training image I_i^u by:

$$LP_i^u = TC_i^u \cdot \exp(-IM_i^u), \tag{4}$$

from which we can observe that the unlabeled training images with higher tag confidence and lower image complexity would be assigned with higher learning reliability values. After each learning stage, we merge as much unlabeled training images as the number of the original labeled training images based on the obtained learning reliability values to form the new labeled training image set for the next learning stage. The obtained learning reliability values of the unlabeled training images essentially form the learning curriculum for guiding the subsequent learning process. Because the learning reliability

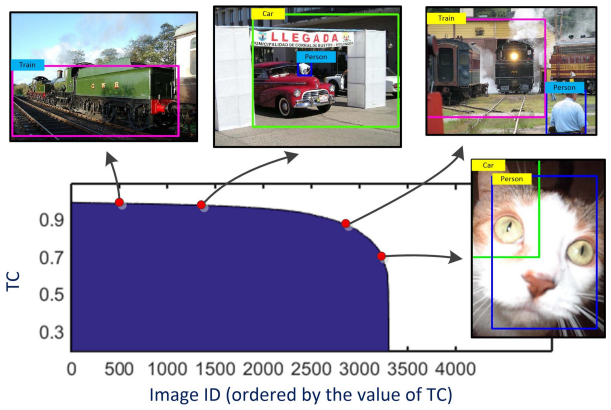


Fig. 5. Examples to illustrate the rationality of building the learning reliability based on the tag confidence (TC), where boxes in different colors indicate the predicted locations of different object categories. The tags in blue are correctly predicted while yellow ones are wrongly predicted. From this figure, we can observe that examples with higher TC values tend to have more reliable prediction of both the object category and location, and thus with higher tag confidence. On the contrary, examples with lower TC values tend to have less reliable prediction of the object category and location, and thus with lower tag confidence.

will be re-inferred after each learning stage, the learning curriculum would be progressively modified and updated, which leads to the progressive learning curriculum applied in the proposed learning framework.

Essentially, our idea to infer the learning reliability of the training samples and gradually involve the confident ones into the learning procedure compiles to the modern self-paced learning (SPL) and curriculum learning (CL) regimes [26]–[32]. Specifically, SPL and CL are two different kinds of weighting-based robust learning regimes. The core idea of the SPL regime is to alternately infer the learning confidence of each training sample and learn the model during the learning stages. While the core idea of the CL regime is to adopt a pre-defined learning curriculum to guide the learning procedure to learn from easy examples to more complex ones. However, the proposed learning framework also has some interesting differences compared with SPL and CL: 1) Instead of inferring the self-paced learning weights via a single learning model, e.g., the support vector machine (SVM) used in [27] and [33], the learning reliability in the proposed framework is established by the Fast R-CNN model and used for training of the WSDDN model, facilitating the effective learning of the proposed cross model co-training process. 2) Rather than pre-defining a learning curriculum and fixing it during the entire learning procedure, the learning curriculum formed by the learning reliability values are established in a progressive manner, i.e., learning reliability values in our framework are modified and updated along the learning stages. Moreover, SPL and CL are usually used under the supervision in some forms of the human annotation, while the proposed learning scheme works in the learning framework with semi-annotated weak labels.

4) Network Co-Training Strategy: For implementing the cross model co-training of the object localizer and the tag generator, three strategies can be used to train these two networks. As shown in Fig. 6 (a), the first network co-training

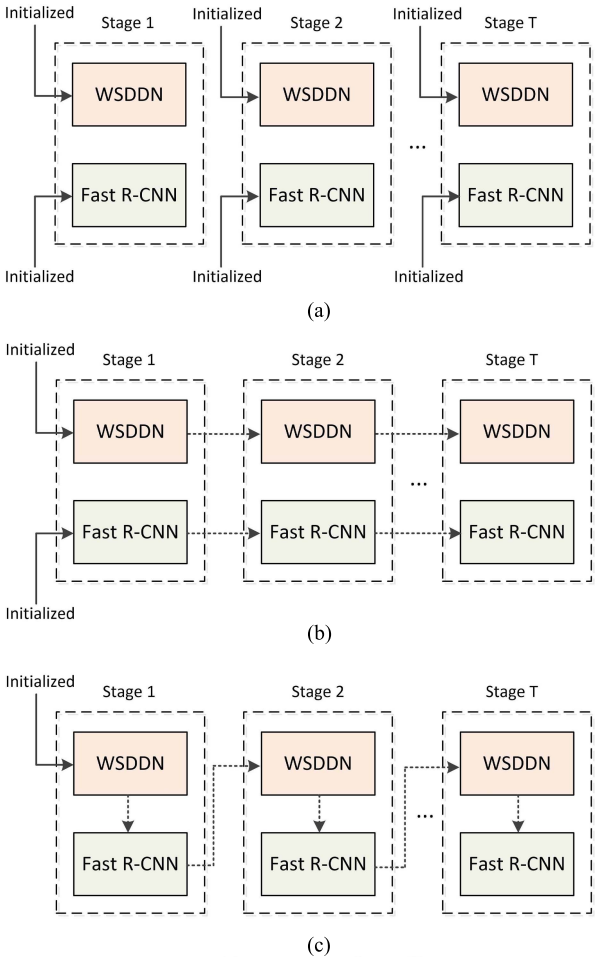


Fig. 6. Brief illustration of the different network co-training strategies. The solid lines and arrows indicate the parameters pre-trained on ImageNet, while the dashed lines and arrows indicate the parameters transferred between different networks or learning stages. (a) The independent initialization strategy. (b) The fine-tuning strategy. (c) The cross-tuning strategy.

strategy is to directly use the training data generated by the previous learning stage to train new object localizer and tag generator (which are initialized by the parameters pre-trained on ImageNet) in the current learning stage. This indicates that the training processes of the object localizer in different learning stages are independent, and so does for the training processes of the tag generator. Thus, we name this strategy as the Independent Initialization strategy, i.e., II for short.

The second network co-training strategy is shown in Fig. 6 (b). In this strategy, the network parameters of both the object localizer and the tag generator are initialized by the parameters pre-trained on ImageNet in the first learning stage. Then, from the second learning stage, the models trained in the previous learning stage would transfer their network parameters to the corresponding model in the current learning stage, which indicates that the object localizer in the current learning stage will be trained by fine-tuning on the network parameters of it trained in the previous learning stage. It goes same for the training of the tag generator. We name this strategy as the Fine-Tuning strategy, i.e., FT for short.

The third network co-training strategy is shown in Fig. 6 (c). In this strategy, the network parameters of the object localizer

are initialized by the parameters pre-trained on ImageNet in the first learning stage. After the training process of the object localizer, the network parameters of its body net will be transferred to initialize the body net parameters of the tag generator. Then, the tag generator will be trained by further fine-tuning on the transferred network parameters. Similarly, after the training process of the tag generator, the network parameters of its body net will be transferred to initialize the body net parameters of the object localizer in the next learning stage. Then, the object localizer will be trained by fine-tuning on these network parameters. This strategy goes along the entire learning stages and we name it as the Cross-Tuning strategy, i.e., CT for short.

In this paper, we use the third Corss-Tuning strategy to train the networks as, intuitively, it can make the two learning models agree on the visual patterns extracted by each of them and thus let them work compatibly and provide the needed informative knowledge to each other.

IV. Experiments

Following the previous works on object localization and detection [18], [20], we evaluate the experimental results of the proposed approach based on two criteria. The first one is the corrected localization rate (CorLoc), which is defined as the percentage of images that contain at least one instance of the target object category for which the most confident instance should be localized correctly. CorLoc is widely used to evaluate the object annotation/localization performance on the weakly labeled training data. The second criterion is the mean average precision (mAP), which is the standard evaluation protocol used to evaluate the object detection performance on the test data. For both criteria, a bounding box is considered to be correctly localized or detected if it has an intersection-over-union ratio of at least 50 percent with a ground-truth object instance. We evaluate our method on three benchmark datasets, which are the Pascal VOC Trainval-2007 [34], Test-2007 [34], and Test-2010 [34], respectively.

For training the object localizer, we run 20 epochs in each stage of the proposed learning procedure. According to [20], the learning rate for the network in the first learning stage was set to 10^{-5} for the first ten epochs and 10^{-6} for the last ten epochs. For the subsequent learning stages, the learning rate for the body net of the object localizer was set as one tenth of the corresponding learning rate used for training the body net of it in the previous learning stage. The learning rate of the rest part of the object localizer was kept the same as the previous learning stage. For training the tag generator, we performed 8 epochs in each learning stage, where the global learning rates were set to be 0.001 and 0.0001 in the first 4 and last 4 epochs, respectively. All layers used a per-layer learning rate of 1 for weights and 2 for biases. Similar to the training of the object localizer, we reduced the learning rate of its body net to one tenth of that used in the previous learning stage.

We followed several previous WSOD approaches [20], [41] to use the Edge Box algorithm [46]. In our experiments, in order to obtain a small set of weakly labeled training images which contain relatively balanced distribution of different object categories, we randomly collected 20 positive training images and 20 negative training images for each object category from the VOC Trainval-2007 (which contains 5011 images). This leads to totally 744 weakly labeled training images because some images may contain more than one object category. Thus, the weakly labeled training data is about 15% of those used in the conventional WSOD approaches. Fig. 7 shows some examples in the obtained training image set, from which we can observe that, albeit small, the collected training image set still contains diverse image scenes, i.e., some of the training images have clear image background and a bit foreground objects, some of other training images have confusing background regions that are similar to the foreground objects, and some of other training images have clustered image contents with a number of objects (with occlusion and interaction) and complex image background. Consistent with some previous works, e.g., [47], the proposed learning procedure typically converges in 3 stages. Thus, we set $T = 3$.

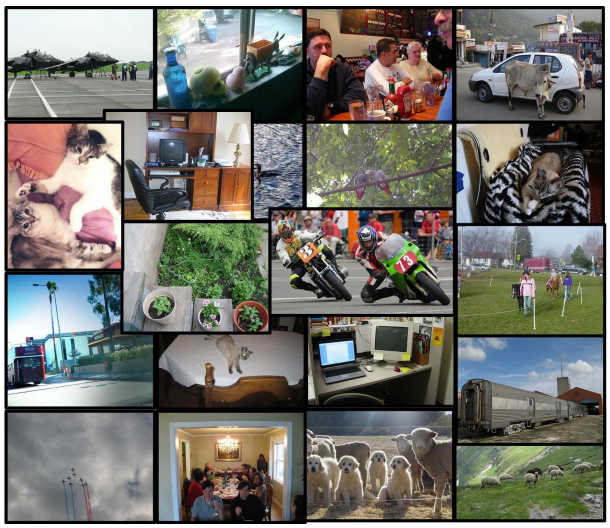


Fig. 7. Some examples of the collected subset of the weakly labeled training images. Learning from such images and weak labels tends to be challenging as, besides the limitation of its scale, some of them contain complex image scenes and multiple objects.

A. Compared With State-of-the-Art WSOD Approaches

In this section, we evaluate the proposed approach by comprehensively comparing it with a number of the existing WSOD approaches in different datasets and under various evaluation criteria. Notice that all the existing WSOD approaches need to learn their object detectors under stronger supervision than the object detectors learnt in the investigated problem of this paper. Specifically, we first compared the proposed approach with the state-of-the-art WSOD approaches, including Zhang *et al.* [14], Siva and Xiang [17], Cinbis *et al.* [18], Bilen and Vedaldi [20], Pandey and Lazebnik [35], Siva *et al.* [36], Shi *et al.* [37], Bilen *et al.* [38], Ren *et al.* [39], Wang *et al.* [40], Singh *et al.* [41], and Li *et al.* [42]. The experiment was implemented on the VOC Trainval-2007 set for the task of object localization, which was

TABLE II

COMPARISON WITH STATE-OF-THE-ART WSOD APPROACHES ON THE VOC TRAINVAL-2007 SET IN TERMS OF CORLOC (HIGHER VALUES INDICATE BETTER PERFORMANCE). NOTICE THAT THE TRAINING PROCESS OF THE PROPOSED APPROACH ONLY LEVERAGES 15% OF THE WEAKLY LABELED TRAINING IMAGES THAT ARE USED IN THESE COMPARED STATE-OF-THE-ARTS

	aero	bike	bird	boat	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pson	plat	shep	sofa	trai	tv	Av
Pandey [35]	50.9	56.7	–	10.6	0.0	56.6	–	–	2.5	–	14.3	–	50.0	53.5	11.2	5.0	–	34.9	33.0	40.6	–
Siva [36]	45.8	21.8	30.9	20.4	5.3	37.6	40.8	51.6	7.0	29.8	27.5	41.3	41.8	47.3	24.1	12.2	28.1	32.8	48.7	9.4	30.2
Siva [17]	42.4	46.5	18.2	8.8	2.9	40.9	73.2	44.8	5.4	30.5	19.0	34.0	48.8	65.3	8.2	9.4	16.7	32.3	54.8	5.5	30.4
Shi [37]	67.3	54.4	34.3	17.8	1.3	46.6	60.7	68.9	2.5	32.4	16.2	58.9	51.5	64.6	18.2	3.1	20.9	34.7	63.4	5.9	36.2
Zhang [14]	71.0	27.2	48.8	40.9	6.6	51.6	46.1	54.6	5.4	58.9	15.5	52.7	60.3	50.6	29.2	17.1	52.1	31.9	56.3	17.6	39.7
Bilen [38]	66.4	59.3	42.7	20.4	21.3	63.4	74.3	59.6	21.1	58.2	14.0	38.5	49.5	60.0	19.8	39.2	41.7	30.1	50.2	44.1	43.7
Ren [39]	79.2	56.9	46.0	12.2	15.7	58.4	71.4	48.6	7.2	69.9	16.7	47.4	44.2	75.5	41.2	39.6	47.4	32.3	49.8	18.6	43.9
Wang [40]	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
Singh [41]	58.8	–	49.6	15.4	–	–	64.9	59.0	–	43.2	–	51.2	57.5	63.1	–	–	–	–	54.4	–	–
Cinbis [18]	65.3	55.0	52.4	48.3	18.2	66.4	77.8	35.6	26.5	67.0	46.9	48.4	70.5	69.1	35.2	35.2	69.6	43.4	64.6	43.7	52.0
Li [42]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
Bilin (L) [20]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
OURS	68.3	58.0	51.1	39.4	24.4	67.0	71.5	61.6	18.2	76.0	22.4	57.7	67.0	76.3	37.1	33.7	64.9	51.6	64.6	63.8	53.7

qualitatively evaluated under the evaluation criteria of CorLoc. The corresponding experimental results are shown in Table II. Surprisingly, the object localization performance obtained by the proposed approach can even outperform all state-of-the-art WSOD approaches. This indicates the novel technique components can effectively offset the loss of the supervision provided by the human annotated image tags. To be specific, compared with Siva and Xiang [17], Cinbis *et al.* [18], Pandey and Lazebnik [35], Siva *et al.* [36], Shi *et al.* [37], and Bilen *et al.* [38] approaches, the proposed approach adopt the more advanced deep learning technique, which can build object detectors based on the learnt richer visual patterns and more powerful feature representations. Compared with Zhang *et al.* [14], Bilen and Vedaldi [20], Ren *et al.* [39], Wang *et al.* [40], Singh *et al.* [41], and Li *et al.* [42] approaches, the proposed approach designed a cross model co-training framework, which can leverage two powerful deep learning models and collaborate them to work compatibly. Such technique is effective and beyond the exploration of most existing WSOD approaches.

For evaluating the object detection performance of the object detectors learnt by the proposed approach, we also compared the proposed approach with several state-of-the-art WSOD approaches, such as Zhang *et al.* [14], Siva and Xiang [17], Cinbis *et al.* [18], Song *et al.* [19], Bilen and Vedaldi [20], Pandey and Lazebnik [35], Bilen *et al.* [38], Ren *et al.* [39], Wang *et al.* [40], Singh *et al.* [41], Li *et al.* [42], Russakovsky *et al.* [43], and Song *et al.* [44] on the VOC Test-2007 set. The object detection results were evaluated by the evaluation criterion of mAP, and the corresponding experimental results are shown in Table III and Table IV. Similar to the object localization performance on the VOC Trainval-2007 set, the proposed approach can outperform the WOSD approaches like Siva and Xiang [17], Song *et al.* [19], Pandey and Lazebnik [35], Bilen *et al.* [38], Russakovsky *et al.* [43], and Song *et al.* [44] mainly due to the adoption of the more powerful deep learning techniques. The proposed approach can also outperform some deep learning-based WSOD approaches like Zhang *et al.* [14], Cinbis *et al.* [18], Ren *et al.* [39], and Wang *et al.* [40], which is probably because of the collaboration scheme for jointly training two deep learning models. Whereas, compared with the most advanced WSOD approaches like Bilen and Vedaldi [20] and Li *et al.* [42], the proposed approach can successfully approach to their performance but cannot outperform them, which indicates the slightly worse generalization capability of the object detectors learnt by the proposed approach. However, for the proposed approach, such obtained performance tends to be already satisfactory and even better than our expectation. Because the proposed approach essentially only uses about 15% of the weakly labeled trailing data used in the state-of-the-art WSOD approaches.

By comparing the results in Table III and Table II, we observe that our approach achieves the best performance in the localization task, while a worse performance in the detection task. One possible explanation is that in the localization task, our model localizes many objects with similar appearance. Although such localization results wont hurt the localization performance in terms of CorLoc, they limit the diversity of the training samples when learning the object detector. This would influence the generalbility of the learnt object detector and thus hurt the detection performance in testing.

In addition, we also compared the proposed approach with the acquirable state-of-the-art WSOD approaches on the VOC Test-2010 set, which are Cinbis *et al.* [18], Bilen and Vedaldi [20], Singh *et al.* [41], and Kantorov *et al.* [45], respectively. On this dataset, we implemented more through experimental comparisons with Bilen and Vedaldi [20] as this approach has the most comparable network architecture with our method, although it requires much more weakly labeled images during its training procedure. Specifically, we com-

TABLE III

COMPARISON WITH STATE-OF-THE-ART WSOD APPROACHES ON THE VOC TEST-2007 SET IN TERMS OF MAP (HIGHER VALUES INDICATE BETTER PERFORMANCE). NOTICE THAT THE TRAINING PROCESS OF THE PROPOSED APPROACH ONLY LEVERAGES 15% OF THE WEAKLY LABELED TRAINING IMAGES THAT ARE USED IN THESE COMPARED STATE-OF-THE-ARTS

	aero	bike	bird	boat	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pson	plat	shep	sofa	trai	tv	Av
Pandey [35]	11.5	–	–	3.0	–	–	–	–	–	–	–	–	20.3	9.1	–	–	–	–	13.2	–	–
Siva [17]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
Russa [43]	30.8	25.0	–	3.6	–	26.0	–	–	–	–	–	–	21.3	29.9	–	–	–	–	–	–	15.0
Song [19]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
Song [44]	36.3	47.6	23.3	12.3	11.1	36.0	46.6	25.4	0.7	23.5	12.5	23.5	27.9	40.9	14.8	19.2	24.2	17.1	37.7	11.6	24.6
Ren [39]	41.3	39.7	22.1	9.5	3.9	41.0	45.0	19.1	1.0	34.0	16.0	21.3	32.5	43.4	21.9	19.7	21.5	22.3	36.0	18.0	25.5
Bilen [38]	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
Zhang [14]	45.4	21.8	35.0	23.8	9.2	50.1	43.0	41.8	1.8	26.9	27.6	37.9	41.2	43.7	17.0	11.9	24.8	22.5	48.8	25.9	30.1
Wang [40]	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
Singh [41]	53.9	–	37.7	13.7	–	–	56.6	51.3	–	24.0	–	38.5	47.9	47.0	–	–	–	–	48.4	–	–
Cinbis [18]	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
Bilen (L) [20]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
Li [42]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	21.8	57.1	22.1	34.4	50.5	61.8	16.2	29.9	40.7	15.9	55.3	40.2	39.5
OURS	50.0	39.8	28.8	11.1	03.1	52.8	55.0	51.0	02.6	47.2	15.4	43.1	50.2	54.1	17.4	10.2	40.5	29.8	44.2	42.1	34.4

TABLE IV

COMPARISON WITH STATE-OF-THE-ART WSOD APPROACHES ON THE VOC TEST-2010 SET IN TERMS OF MAP (HIGHER VALUES INDICATE BETTER PERFORMANCE). NOTICE THAT THE TRAINING PROCESS OF THE PROPOSED APPROACH ONLY LEVERAGES 15% OF THE WEAKLY LABELED TRAINING IMAGES THAT ARE USED IN THESE COMPARED STATE-OF-THE-ARTS

	aero	bike	bird	boat	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pson	plat	shep	sofa	trai	tv	Av
Singh [41]	53.5	–	37.5	8.0	–	–	44.2	49.4	–	33.7	–	43.8	42.5	47.6	–	–	–	–	40.6	–	–
Cinbis [18]	44.6	42.3	25.5	14.1	11.0	44.1	36.3	23.2	12.2	26.1	14.0	29.2	36.0	54.3	20.7	12.4	26.5	20.3	31.2	23.7	27.4
Bilen's(F) [20]	54.0	54.1	35.8	8.8	19.7	53.7	37.8	34.2	8.6	29.3	6.7	39.0	47.7	62.6	12.1	18.2	29.5	23.1	46.3	38.0	33.0
Bilen (L) [20]	66.3	33.2	47.2	25.3	6.3	47.7	30.1	61.7	5.2	43.2	2.5	61.7	50.7	51.5	21.1	4.8	30.5	13.9	52.8	16.0	33.6
Bilen's(M) [20]	54.3	47.7	34.5	18.4	20.7	55.2	40.7	31.6	10.2	34.5	17.6	27.3	44.3	61.1	9.4	19.8	32.0	28.1	48.1	40.7	33.8
Kantorov's [45]	63.4	53.1	34.7	6.0	15.6	52.3	43.2	33.0	6.1	34.3	38.2	47.7	44.2	60.5	21.1	16.6	24.1	33.0	32.1	46.3	35.3
OURS	55.1	41.6	25.0	05.3	02.7	51.9	40.1	51.8	03.5	38.0	03.4	43.5	43.6	52.5	15.2	07.7	39.6	22.7	34.0	36.1	30.7

pared with three settings of Bilen and Vedaldi [20], which are the Bilen (L), Bilen (M), and Bilen (F), respectively. The experimental results reported in Table IV show that the proposed approach can not only outperform Cinbis *et al.* [18] and Singh *et al.* [41], but also approach to Bilen and Vedaldi [20] and Kantorov *et al.* [45], even only using about 15% of the training data that are used in these state-of-the-art methods.

By running on a workstation with two 2.1 GHz 8-core CPUs, 252 GB memory, additionally with a GTX 1080Ti GPU, it takes nearly 71 hours to train our model and 0.14 seconds to perform the testing process per image. The training time of our approach is reasonably longer than the conventional weakly supervised object detection method, e.g., WSDDN which takes about 10 hours to train, as it requires multiple stages of training. As for the testing speed, our approach is faster than WSDDN which needs 2 seconds for performing each image.

B. Model Analysis

In this section, we comprehensively analyzed the properties of the proposed learning model. Firstly, we implemented the abolition study to evaluate some of the considered factors in the proposed approach. To be specific, we compared the proposed approach with the following five baseline approaches:

- **Int WSDDN**: The first baseline is used to evaluate the performance of one state-of-the-art WSOD approach,[3] i.e., the WSDDN model, which has the same network architecture with the object localizer as in our framework. In this baseline, we directly trained WSDDN on the small set of weakly labeled training images.
- **MS WSDDN**: This baseline is established by only using a object localizer, i.e., the WSDDN model, in an multi-stage learning procedure. It first trained the initial WSDDN based on the small set of weakly labeled training images. Then, the learnt WSDDN model was used to predict the image tag for each unlabeled training images. We also generate the progressive learning curriculum to select confident training samples after each training stage. The whole learning procedure conducts three-stage training of the WSDDN model.
- **OURS w/o LC**: This baseline is to implement the proposed cross model co-training process without using

[3]All the experimental results of the WSDDN approach reported in the following parts are obtained based on the 'WSDDN-L' model of [20].

TABLE V

COMPARISON RESULTS BETWEEN THE PROPOSED APPROACH AND THE FIVE BASELINE METHODS ON THE VOC TRAINVAL-2007 SET IN TERMS OF CORLOC (HIGHER VALUES INDICATE BETTER PERFORMANCE)

	aero	bike	bird	boat	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pson	plat	shep	sofa	trai	tv	Av
Int WSDDN	57.1	48.6	37.8	31.9	24.4	56.9	58.1	29.9	19.1	58.2	18.3	28.4	57.8	66.3	18.6	32.6	63.9	36.6	52.9	54.1	42.6
MS WSDDN	53.3	52.2	40.2	34.6	24.0	60.4	58.1	32.6	17.3	58.9	20.5	29.8	59.9	66.7	18.1	30.0	69.1	34.7	54.8	54.1	43.5
OURS w/o CL	68.8	59.2	56.8	41.0	32.8	70.6	66.5	59.3	21.3	71.9	33.1	49.1	53.4	77.5	30.4	31.9	68.0	49.5	65.0	56.6	53.1
OURS w II	62.5	49.0	55.6	41.5	28.6	64.0	69.0	47.7	32.0	71.2	17.5	45.3	63.9	74.3	30.1	31.9	68.0	53.5	66.5	72.0	52.2
OURS w FT	58.3	45.9	56.5	36.2	28.6	66.0	68.6	63.4	17.7	71.9	11.8	55.3	64.6	73.1	33.8	28.9	73.2	53.0	68.4	65.9	52.1
OURS (w CT)	68.3	58.0	51.1	39.4	24.4	67.0	71.5	61.6	18.2	76.0	22.4	57.7	67.0	76.3	37.1	33.7	64.9	51.6	64.6	63.8	53.7

TABLE VI

COMPARISON RESULTS BETWEEN THE PROPOSED APPROACH AND THE FIVE BASELINE METHODS ON THE VOC TEST-2007 SET IN TERMS OF MAP (HIGHER VALUES INDICATE BETTER PERFORMANCE)

	aero	bike	bird	boat	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pson	plat	shep	sofa	trai	tv	Av
Int WSDDN	41.8	26.0	21.1	15.0	09.2	45.1	36.1	18.7	04.8	28.5	05.5	13.9	38.3	45.4	06.4	08.9	23.2	25.8	35.2	43.1	24.6
MS WSDDN	38.8	32.4	21.6	14.7	06.3	50.5	39.8	18.7	04.1	29.3	13.8	18.0	38.5	41.3	06.8	06.9	23.3	20.9	35.7	40.9	25.1
OURS w/o CL	53.8	42.5	33.4	15.6	13.0	55.1	52.4	54.0	03.9	40.7	20.2	40.0	26.5	50.4	12.3	09.9	30.2	35.5	41.9	41.3	33.6
OURS w II	48.1	38.6	26.8	17.9	05.7	54.9	52.0	36.7	07.8	48.1	16.1	24.2	38.1	48.8	12.6	13.0	40.4	38.0	48.4	49.3	33.3
OURS w FT	43.7	35.0	26.1	14.5	10.9	55.6	52.1	52.8	04.6	44.8	05.2	44.3	47.9	48.0	16.0	11.1	43.6	30.0	45.9	50.0	34.1
OURS (w CT)	50.0	39.8	28.8	11.1	03.1	52.8	55.0	51.0	02.6	47.2	15.4	43.1	50.2	54.1	17.4	10.2	40.5	29.8	44.2	42.1	34.4

the learning curriculum. Specifically, in this baseline, all the unlabeled training images, together with their corresponding image tags generated in the previous learning stage, were used to train the WSDDN model in the current learning stage.

- **OURS w II**: This baseline indicates to adopt the Independent Initialization strategy to train the network parameters of the WSDDN model and the Fast R-CNN model, instead of the Corss-Tuning strategy. More detailed descriptions about the Independent Initialization strategy can be referred to in Sec. III-B.
- **OURS w FT**: This baseline indicates to adopt the Fine-Tuning strategy to train the network parameters of the WSDDN model and the Fast R-CNN model, instead of the Corss-Tuning strategy. More detailed descriptions about the Fine-Tuning strategy can be referred to in Sec. III-B.

The comparison results between the proposed approach and the aforementioned baseline methods are reported in Table. V and Table. VI. Specifically, Table. V shows the comparison results on Trainval-2007 in terms of the CorLoc, while Table. VI shows the comparison results on Test-2007 in terms of the mAP. From these two tables, we can observe that: 1) The experimental results of the **Int WSDDN** baseline indicates that reducing the weakly labeled training images of the WSDDN model would cause obvious but rationale performance drop, and directly using the state-of-the-art WSOD approach cannot obtain satisfactory performance in the investigated problem. 2) The experimental results of the **MS WSDDN** baseline indicates that only using a object localizer in the multi-stage learning procedure cannot obtain as strong object detectors as training the object localizer and the tag generator jointly. This demonstrates the insight of the paper, i.e., jointly training the function-distinct deep models in an effective cross model co-training framework can make them work collaboratively and boost the performance of the learnt object detectors. 3) Comparison between the proposed approach, i.e., **OURS (w CT)**, and the **OURS w/o LC** demonstrates the effectiveness of the established learning curriculum for guiding the cross model co-training process. 4) Comparison among the proposed approach, i.e., **OURS (w CT)**, with the **OURS w II** baseline and the **OURS w FT** indicates the better performance of the adopted Corss-Tuning strategy in the proposed cross-model co-training framework.

Next, we verified the performance improvement between each learning stage. This experiment was implemented by testing the object detectors trained on the collected subset of VOC Trainval-2007 dataset (20 positive training images and 20 negative training images per object category, totally 744 images) after each learning stage. The testing process was performed on the Test-2007 dataset. The corresponding results are reported in the top block of Table VII, from which we can observe that the proposed cross-model co-training framework can guide a steady performance improvement after each learning stage. To illustrate the improvement better, we also visualize some improved cases and failure cases in Fig. 8.

Finally, we implemented the experiments to study how the performance of the proposed approach changes when increasing the amount of weakly labeled training images. Specifically, we implemented our approach (with three learning stages) under two additional settings. The first experimental setting (**OURS@50**) is to collect 50 positive training images and 50 negative training images for each object category to form

TABLE VII

EVALUATION RESULTS IN DIFFERENT LEARNING STAGES AND UNDER DIFFERENT EXPERIMENTAL SETTINGS ON THE VOC TEST-2007 SET IN TERMS OF MAP (HIGHER VALUES INDICATE BETTER PERFORMANCE)

	aero	bike	bird	boat	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pson	plat	shep	sofa	trai	tv	Av
OURS@20																					
-stage 1	46.7	37.0	26.7	18.2	06.2	48.9	49.5	32.4	03.3	46.1	22.1	26.3	37.0	49.9	04.9	15.5	36.6	30.5	42.6	45.4	31.3
-stage 2	48.1	38.6	26.8	17.9	05.7	54.9	52.0	36.7	07.8	48.1	16.1	24.2	38.1	48.8	12.6	13.0	40.4	38.0	48.4	49.3	33.3
-stage 3	50.0	39.8	28.8	11.1	03.1	52.8	55.0	51.0	02.6	47.2	15.4	43.1	50.2	54.1	17.4	10.2	40.5	29.8	44.2	42.1	34.4
OURS@50	46.7	38.8	29.6	27.7	15.0	60.8	57.0	52.5	02.6	45.8	21.3	26.8	48.7	56.0	12.1	13.0	39.4	37.6	46.0	56.8	36.7
WSDDN@50	37.9	29.4	23.0	21.6	09.1	60.3	42.9	23.1	04.2	30.5	14.4	14.0	32.8	46.2	05.4	08.2	29.2	22.4	46.4	47.8	27.4
OURS@100	50.7	41.0	35.1	25.3	14.0	64.3	57.3	41.2	08.8	53.7	22.2	34.1	55.4	55.0	11.7	14.7	50.2	40.2	40.1	49.7	38.2
WSDDN@100	43.8	38.3	26.3	21.0	11.1	61.6	47.1	26.2	08.6	38.7	12.4	22.2	32.7	54.2	04.3	10.4	37.3	32.8	50.3	50.8	31.5

Fig. 8. Some visualization examples of the performance improvement obtained by the iterative learning process. Successful cases are shown in top two rows while failure cases are shown in bottom two rows. To our best knowledge, the failure cases might be caused by the semantically discriminative object parts, poor light condition, or clutter background.

the weakly labeled training set with 1676 images (about 33% of the entire training set) and use it to train the object detectors. The second experimental setting (**OURS@100**) is to collect 100 positive training images and 100 negative training images for each object category to form the weakly labeled training set with 2904 images (about 58% of the entire training set) and use it to train the object detectors.

In the bottom block of Table VII, we report the experimental results under these two experimental settings on Test-2007 in terms of the mAP. Besides **OURS@50** and **OURS@100**, we also report **WSDDN@50** and **WSDDN@100** to indicate the experimental results of the baseline approach WSDDN under these experimental settings, respectively. From the experimental results, we can observe that the proposed approach can consistently obtain significant performance boost as compared with the corresponding baseline under these experimental settings. Specifically, when collecting 50 positive training images and 50 negative training images per object category to form the weakly labeled training set, the proposed approach can outperform the corresponding baseline by around 9.3% in terms of mAP score. When collecting 100 positive training images and 100 negative training images per object category to form the weakly labeled training set, the proposed approach can outperform the corresponding baseline by around 6.7% in terms of mAP score. With such consistent performance gain, the proposed approach is demonstrated to have the potential to be scaled up for further improving the performance of the existing WSOD approaches.

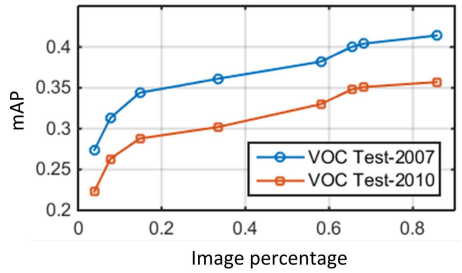


Fig. 9. The mAP v.s. image percentage curves on the VOC Test-2007 Set and VOC Test-2010 Set.

For further studying the influence of using different percentage of the weakly labeled training images, we conduct more comprehensive experiments and draw the obtained results in the mAP v.s. image percentage curves in Fig. 9. From Fig. 9 we can observe that 1) When using less than 15 percentage of weakly supervised training images, the performance of our approach drops dramatically. 2) Adding more weakly labeled training images into our learning approach leads to a consistent performance gain on both datasets. 3) When using 68 and 85 percentage of weakly labeled training images, our approach starts to outperform the best state-of-the-art method on the VOC Test-2007 and VOC Test-2010, respectively.

To evaluate how much further the performance of our approach goes up when leveraging a larger image dataset, we implement an additional experiment, named **OURS@50+**, to add the images from the VOC Trainval-2012 set[4] (without their annotations) into the training process of our approach under the setting of **OURS@50**. The only difference between **OURS@50+** and **OURS@50** is the amount of unlabeled training images. According to our experiment, **OURS@50+** obtains 48.1 mAP on the VOC Test-2007 Set, which outper-

[4]The images contained by both VOC Trainval-2012 and VOC Trainval-2007 have been screened.

forms **OURS@50** by more than 11 percent. This experiment demonstrates that our approach has a good scale up capacity by using more unlabeled training images.

V. Conclusion

In this paper, we have made a further step to alleviate the human labor for training object detectors, which leads to the investigation of the novel problem of learning object detectors with semi-annotated weak labels. For solving this problem, we have proposed a brand new cross model co-training framework, which can jointly train two powerful deep learning models, i.e., the object localizer and the tag generator, in a leaning curriculum-guided multi-stage learning procedure. With beneficial information (visual patterns and supervision labels) transferred between these two models, the learning framework can progressively improve each of the deep models and finally obtain strong object detectors. Comprehensive experiments on three benchmark datasets have been conducted to demonstrate the effectiveness of the proposed approach. Notably, by only using about 15% of the weakly labeled training images that are used in the existing WSOD method, the proposed approach can effectively approach to, or even outperform, the state-of-the-art WSOD approaches. For future work, we plan to mine the co-salient patterns [27], [48]–[51] from the weakly labeled training images to further improve the performance of ODSAWL.

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[2] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *Proc. ICCV*, Nov. 2011, pp. 89–96.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[4] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[5] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

[6] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Dec. 2015, pp. 1440–1448.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, Jun. 2016, pp. 779–788.

[8] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[9] I. Misra, A. Shrivastava, and M. Hebert, "Watch and learn: Semi-supervised learning of object detectors from videos," in *Proc. CVPR*, Jun. 2015, pp. 3593–3602.

[10] Y.-X. Wang and M. Hebert, "Model recommendation: Generating object detectors from few samples," in *Proc. CVPR*, Jun. 2015, pp. 1619–1628.

[11] M. H. Nguyen, L. Torresani, F. De La Torre, and C. Rother, "Weakly supervised discriminative localization and classification: A joint learning process," in *Proc. ICCV*, Sep./Oct. 2009, pp. 1925–1932.

[12] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[13] Z. Yan, J. Liang, W. Pan, J. Li, and C. Zhang. (2017). "Weakly- and semi-supervised object detection with expectation-maximization algorithm." [Online]. Available: https://arxiv.org/abs/1702.08740

[14] D. Zhang, D. Meng, L. Zhao, and J. Han, "Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning," in *Proc. IJCAI*, 2016, pp. 3538–3544.

[15] Y. Yang, G. Shu, and M. Shah, "Semi-supervised learning of feature hierarchies for object detection in a video," in *Proc. CVPR*, Jun. 2013, pp. 1650–1657.

[16] Y. Tang, J. Wang, B. Gao, E. Dellandréa, R. Gaizauskas, and L. Chen, "Large scale semi-supervised object detection using visual and semantic knowledge transfer," in *Proc. CVPR*, Jun. 2016, pp. 2119–2128.

[17] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *Proc. ICCV*, Nov. 2011, pp. 343–350.

[18] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.

[19] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. (2014). "On learning to localize objects with minimal supervision." [Online]. Available: https://arxiv.org/abs/1403.1024

[20] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. CVPR*, Jun. 2016, pp. 2846–2854.

[21] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan, "Towards computational baby learning: A weakly-supervised approach for object detection," in *Proc. ICCV*, Dec. 2015, pp. 999–1007.

[22] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. (2017). "Deep self-taught learning for weakly supervised object localization." [Online]. Available: https://arxiv.org/abs/1704.05188

[23] Z.-H. Zhou and M. Li, "Semi-supervised regression with co-training," in *Proc. IJCAI*, 2005, pp. 908–913.

[24] A. Levin, P. Viola, and Y. Freund, "Unsupervised improvement of visual detectors using cotraining," in *Proc. ICCV*, Oct. 2003, pp. 626–633.

[25] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. ACM-COLT*, 1998, pp. 92–100.

[26] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. NIPS*, 2010, pp. 1189–1197.

[27] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[28] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3249–3260, Jul. 2016.

[29] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML*, 2009, pp. 41–48.

[30] D. Zhang, L. Yang, D. Meng, D. Xu, and J. Han, "SPFTN: A self-paced fine-tuning network for segmenting objects in weakly labelled videos," in *Proc. CVPR*, Jul. 2017, pp. 5340–5348.

[31] D. Zhang, J. Han, Y. Yang, and D. Huang, "Learning category-specific 3D shape models from weakly labeled 2D images," in *Proc. CVPR*, Jul. 2017, pp. 3587–3595.

[32] L. Han *et al.*, "Self-paced mixture of regressions," in *Proc. IJCAI*, 2017, pp. 1816–1822.

[33] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *Proc. NIPS*, 2014, pp. 2078–2086.

[34] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[35] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Proc. ICCV*, Nov. 2011, pp. 1307–1314.

[36] P. Siva, C. Russell, and T. Xiang, "In defence of negative mining for annotating weakly labelled data," in *Proc. ECCV*, 2012, pp. 594–608.

[37] Z. Shi, T. M. Hospedales, and T. Xiang, "Bayesian joint modelling for object localisation in weakly labelled images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1959–1972, Oct. 2015.

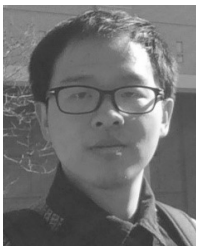
[38] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *Proc. CVPR*, Jun. 2015, pp. 1081–1089.

[39] W. Ren, K. Huang, D. Tao, and T. Tan, "Weakly supervised large scale object localization with multiple instance learning and bag splitting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 405–416, Feb. 2016.

[40] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *Proc. ECCV*, 2014, pp. 431–445.

[41] K. K. Singh, F. Xiao, and Y. J. Lee, "Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection," in *Proc. CVPR*, Jun. 2016, pp. 3548–3556.

[42] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. CVPR*, Jun. 2016, pp. 3512–3520.
[43] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," in *Proc. ECCV*, 2012, pp. 1–15.
[44] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell, "Weakly-supervised discovery of visual pattern configurations," in *Proc. NIPS*, 2014, pp. 1637–1645.
[45] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in *Proc. ECCV*, 2016, pp. 350–365.
[46] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. ECCV*, 2014, pp. 391–405.
[47] Y. Li, M. Paluri, J. M. Rehg, and P. Dollár, "Unsupervised learning of edges," in *Proc. CVPR*, Jun. 2016, pp. 1619–1627.
[48] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *Trans. Intell. Syst. Technol.*, vol. 9, no. 4, 2018, Art. no. 38.
[49] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, Oct. 2018.
[50] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1639–1651, Apr. 2018.
[51] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.



Dingwen Zhang received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2018. From 2015 to 2017, he was a Visiting Scholar with the Robotic Institute, Carnegie Mellon University. He is currently an Associate Professor with the School of Mechano-Electronic Engineering, Xidian University. His research interests include computer vision and multimedia processing, especially on saliency detection, video object segmentation, and weakly supervised learning.

Junwei Han (M'12–SM'15) received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2003. He was a Research Fellow with Nanyang Technological University, The Chinese University of Hong Kong, and the University of Dundee. He is currently a Professor with Northwestern Polytechnical University. His research interests include computer vision and brain imaging analysis. He has published over 100 papers in IEEE Transactions and top tier conferences. He is currently an Associate Editor of the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, *Neurocomputing*, and *Machine Vision and Applications*.

Guangyu Guo received the B.E. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Automation, Northwestern Polytechnical University. His research interests include computer vision and multimedia processing, especially on object detection and deep learning.

Long Zhao received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2016, where he is currently pursuing the M.S. degree with the School of Automation. His research interests include computer vision and multimedia processing, especially on weakly supervised learning and deep learning.