Boosting Weakly Supervised Object Detection via Learning Bounding Box Adjusters

Bowen Dong[1] Zitong Huang[1] Yuelin Guo[1] Qilong Wang[2] Zhenxing Niu[3] Wangmeng Zuo[1,4]✉
[1]Harbin Institute of Technology [2]Tianjin University [3]Alibaba Damo Academay [4]Pazhou Lab
{cndongsky, zitonghuang99, zhenxingniu}@gmail.com gyl2565309278@163.com qlwang@tju.edu.cn wmzuo@hit.edu.cn

Abstract

Weakly-supervised object detection (WSOD) has emerged as an inspiring recent topic to avoid expensive instance-level object annotations. However, the bounding boxes of most existing WSOD methods are mainly determined by precomputed proposals, thereby being limited in precise object localization. In this paper, we defend the problem setting for improving localization performance by leveraging the bounding box regression knowledge from a well-annotated auxiliary dataset. First, we use the well-annotated auxiliary dataset to explore a series of learnable bounding box adjusters (LBBAs) in a multi-stage training manner, which is class-agnostic. Then, only LBBAs and a weakly-annotated dataset with non-overlapped classes are used for training LBBA-boosted WSOD. As such, our LBBAs are practically more convenient and economical to implement while avoiding the leakage of the auxiliary well-annotated dataset. In particular, we formulate learning bounding box adjusters as a bi-level optimization problem and suggest an EM-like multi-stage training algorithm. Then, a multi-stage scheme is further presented for LBBA-boosted WSOD. Additionally, a masking strategy is adopted to improve proposal classification. Experimental results verify the effectiveness of our method. Our method performs favorably against state-of-the-art WSOD methods and knowledge transfer model with similar problem setting. Code is publicly available at `https://github.com/DongSky/lbba_boosted_wsod`.

1. Introduction

Object detection [8, 7, 19, 16] has attracted considerable attention in computer vision community, and benefits a wide range of applications. Along with the development of powerful convolutional neural networks (CNNs) and large-scale well-annotated datasets, the performance of object detection networks has achieved remarkable improvement. Nevertheless, the success of object detection networks highly depends on precise but costly instance-level bounding box annotations of abundant images. To alleviate this issue, weakly supervised object detection (WSOD) aiming at learning effective detection models with image-level supervision has emerged as an inspiring recent topic.

Existing WSOD methods [3, 25, 34, 20] usually adopt the multiple instance learning (MIL) framework based on the precomputed proposals. And most efforts have been given to improve proposal classification ability. However, the bounding boxes of most existing methods are mainly determined by precomputed proposals, thereby being limited in precise object localization. For single-phase WSOD methods [3, 26, 25, 22, 13], the precomputed proposals classified to a specific class are directly taken as the detection results. Bounding box regression branches are introduced in [33, 20, 34] and multi-phase training are adopted in [36, 2]. But they are usually supervised based on the pseudo ground-truths by selecting precomputed proposals with the highest scores. In terms of localization performance, there remains a huge gap between WSOD methods and their fully-supervised counterparts.

Transfer learning has also been investigated to improve the localization performance of WSOD. Lee *et al*. [12] presented a universal bounding box regressor (UBBR) trained on a well-annotated auxiliary dataset for refining bounding boxes generated in WSOD. Instead, Uijlings *et al*. [28] trained a universal detector on the well-annotated source dataset, which is then transferred to WSOD as a generic proposal generator. However, [12] and [28] adopt the single-stage transfer strategy, which actually are not specified to WSOD [3, 26, 12, 28] and suffer from imperfect annotations in source domain [17, 6, 28]. Going beyond [28], Zhong *et al*. [37] trained and exploited the one-class universal detector (OCUD) in a progressive manner. In contrast, both the source well-annotated and target weakly annotated datasets are required in the whole training process for OCUD [37]. When the source dataset is private and is of large scale [24, 18], it is preferred to avoid the direct joint use of the source and target datasets for WSOD with transfer learning. Instead, the owner of source datasets can first extract knowledge from data and then distribute knowledge instead of source datasets to the user for boosting WSOD.

In this paper, we follow the problem setting in [12, 28], and propose a learnable bounding box adjuster (LBBA) for boosting WSOD performance. Specifically, we consider a well-annotated auxiliary dataset and a weakly annotated dataset. Our method involves two subtasks, *i.e.*, learning class-agnostic bounding box adjuster and training LBBA-boosted WSOD model. In comparison to [12, 28], the LBBAs are specifically designed for improving WSOD performance by developing a multi-stage scheme. Different from [37], only the LBBAs and weakly-annotated dataset are used for boosting WSOD, and thus our approach is practically convenient and economical for WSOD training while avoiding the leakage of the auxiliary dataset.

To better learn LBBAs from the well-annotated auxiliary dataset and exploit them to improve the performance of WSOD, we formulate the learning of LBBAs as a bi-level optimization problem and present an EM-like multi-stage training algorithm. In particular, the lower subproblem is formulated to learn a deep detection model by incorporating WSOD with LBBA-based regularization, while the upper subproblem is formulated to learn the boundary box adjuster for regressing the selected region proposals generated by WSOD towards the ground-truth bounding boxes. With such formulation, the LBBAs can thus be learned for optimizing WSOD performance. For solving the bi-level optimization problem, we adopt an EM-like multi-stage training algorithm by alternating between training LBBA and WSOD models. Given the class-agnostic and multi-stage LBBAs, the training of LBBA-boosted WSOD also involves several stages. In each stage, the final LBBA can be used to predict the bounding boxes based on the selected region proposals generated by WSOD, which are then used to train the WSOD models.

Nevertheless, our LBBAs improve localization performance but are limited in improving proposal classification. As a remedy, we introduce a masking strategy to improve the classification performance of the detector. Specifically, a multi-label classifier is introduced to predict category confidence on image-level, which can further suppress scores of false-positive proposals of WSOD network.

Extensive experiments have been conducted to evaluate our proposed method. Benefiting from the class-agnostic setting, LBBAs generalize well to new classes of objects and improves the localization performance of WSOD. Our method performs favorably against state-of-the-art WSOD methods as well as knowledge transfer models with similar problem setting, *e.g.*, UBBR [12]. Contributions of this work can be summarized as follows:

1) Multi-stage learnable bounding box adjusters are presented for improving localization performance of WSOD, which is the core component of our proposed framework. Particularly, LBBAs make it feasible to use source and target datasets separately for training WSOD models, which is practically more convenient and economical.
2) A bi-level optimization formulation, as well as an EM-like multi-stage training algorithm, are suggested to learn LBBAs specified for optimizing WSOD.
3) An effective masking strategy is introduced to improve the accuracy of the proposal classification branch.
4) Experimental results show our proposed method performs favorably against the state-of-the-art WSOD methods and knowledge transfer models with the similar problem setting.

2. Related Work

2.1. Weakly Supervised Object Detection

Weakly supervised object detection (WSOD) aims at training an effective detector only using image-level labels, and is usually formulated as a multiple instance learning (MIL) problem [5]. Existing WSOD approaches can be roughly grouped into two categories: single-phase training methods and multi-phase training ones. For single-phase training methods, they rely on precomputed proposals [29, 1, 38] during training and testing. Specifically, Bilen *et al.* [3] proposed a two-stream detection network (WSDDN) as the basic proposal classifier. To improve *proposal classification ability*, OICR [26] and PCL [25] proposed online classifier refinement module. OIM [15] proposed spatial and appearance graphs with object instance reweighted loss to resolve part domination. SDCN [13] and WS-JDS [22] introduced segmentation branch and collaboration loop to reweight proposals. As for improving *proposal localization ability*, Yang *et al.* [33], WSOD2 [34] and MIST [20] introduced bounding box regression into WSOD network, where proposals with highest scores are selected as pseudo ground-truths to supervise bounding box regression branch.

For multi-phase training methods [36, 35, 13, 30, 32], an additional detector is further trained by selecting proposals with the highest scores as pseudo ground-truths based on the output of trained WSOD network in the prior phase [7]. Any single-phase methods [26, 25, 33, 2] can be extended to multi-phase setting by this procedure. Current multi-phase training methods focus on how to select pseudo ground-truths with the highest scores. However, these approaches rely on only selected precomputed proposals to localize objects or supervise box regression branch, low precision proposals restrict the localization ability of WSOD approaches. Different from the above methods, we aim at resolving this issue by using learnable bounding box adjusters, which provide more precise pseudo boxes supervision to help WSOD network obtain better object localization ability.

2.2. Transfer Learning in WSOD

Transfer learning based WSOD usually leverages an auxiliary dataset to provide semantic information or class-

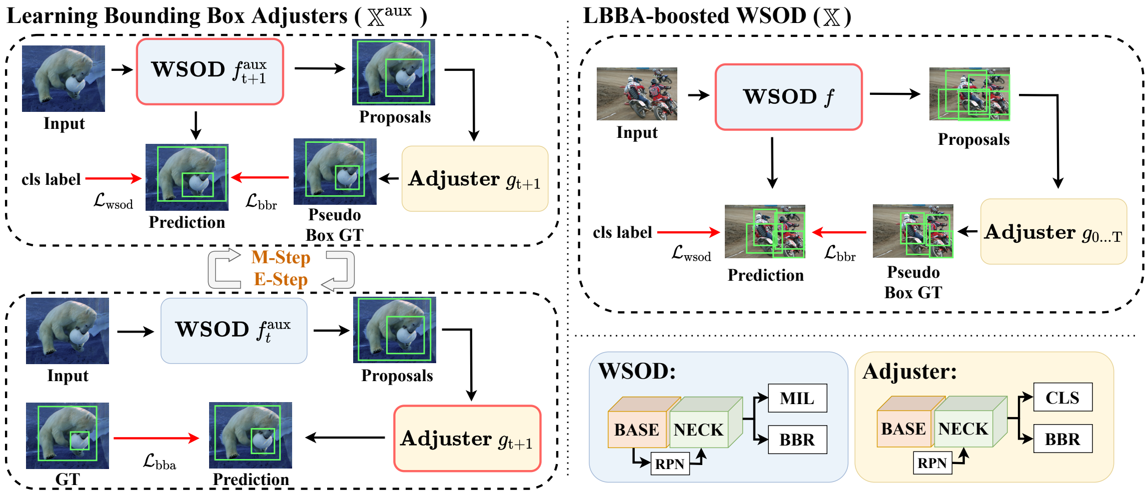

Figure 1. Illustration of our proposed method which includes two subtasks, *i.e.*, **learning bounding box adjusters** (left) and **LBBA-boosted WSOD** (right). For learning bounding box adjusters, we adopt an EM-like algorithm. In **E-step**, adjuster g predicts bounding boxes from proposals of f^{aux} and supervised by ground-truths of $\mathbb{X}^{\text{aux}}$; In **M-step**, WSOD network f^{aux} is supervised by image label as well as adjusted boxes from g on $\mathbb{X}^{\text{aux}}$. For LBBA-boosted WSOD, WSOD network f is supervised by image label and adjusted boxes from g on $\mathbb{X}$. Finally, the learned f is used for evaluation.

agnostic information to help WSOD networks train on weakly-annotated target dataset. Previous works [9, 11, 27] focused on *transferring semantic information* between strong classifier and weakly supervised detector. Among them, Hoffman *et al.* [11] proposed LSDA, which introduces category specific adaptation to adapt a classifier into target detection dataset. Tang *et al.* [27] further extended LSDA by building visual similarity and semantic relatedness. Nonetheless, above methods are not proposed for improving bounding box regression.

Recently, several approaches [21, 14, 28, 12, 37] have been studied to exploit transfer learning for *improving object localization performance*. [21, 14, 28, 37] proposed to learn proposal generators to help WSOD network locate novel objects on weakly-annotated target dataset. Among them, [21, 14, 28] trained proposal generators merely using the auxiliary dataset, while Zhong *et al.* trained generator on both auxiliary dataset and weakly-annotated dataset progressively to generalize better on target dataset. Instead, Lee *et al.* [12] proposed a box refinement module, which takes the random transformations of ground-truth boxes as the input to learn class-agnostic box regressor, and also exhibits certain generalization ability on target weakly-annotated dataset. However, the real boxes generated during WSOD training may be quite different from those by random transformations, making the learned regressor not tailored to WSOD. In comparison to existing methods, our LBBAs can be considered as the multi-stage training of box refinement modules only using the auxiliary dataset, and achieves very competitive box regression performance on weakly-annotated dataset. Different from UBBR[12], our method dynamically takes the proposals generated by WSOD as the input to train LBBA, and thus is expected to achieve improved detection performance.

3. Proposed Method

3.1. Problem Setting and Notations

In this work, we follow the problem setting in [21, 14, 28, 12] for WSOD by using a well-annotated auxiliary dataset $\mathbb{X}^{\text{aux}}$ and a weakly annotated dataset $\mathbb{X}$. In particular, $\mathbb{X}^{\text{aux}}$ is first used to train class-agnostic learnable bounding box adjusters (LBBAs). Then, we utilize both LBBAs and any weakly annotated dataset $\mathbb{X}$ to learn a better WSOD model. For the image-level weakly annotated dataset $\mathbb{X} = \{\mathbf{I}, \mathbb{P}, \mathbf{y}\}$, $\mathbf{I}$ denotes an image from $\mathbb{X}$, and $\mathbf{y}$ denotes the corresponding image-level labels. For the end of WSOD, MCG [1] and selective search [29] are used to extract a set of precomputed proposals $\mathbb{P} = \{\mathbf{p}\}$ for each image $\mathbf{I}$. Besides $\mathbb{X}$, we also introduce a well-annotated auxiliary dataset $\mathbb{X}^{\text{aux}} = \{(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}, \{\mathbf{b}^{\text{aux}}\}, \mathbf{y}^{\text{aux}})\}$. For an image $\mathbf{I}^{\text{aux}}$ from $\mathbb{X}^{\text{aux}}$, $\mathbf{y}^{\text{aux}}$ denotes the image-level labels, and $\{\mathbf{b}^{\text{aux}}\}$ denotes the annotated bounding boxes. To aid WSOD, we also give the precomputed proposals $\mathbb{P}^{\text{aux}} = \{\mathbf{p}^{\text{aux}}\}$ of $\mathbf{I}^{\text{aux}}$. To show the generalization ability of LBBA, we assume the object classes in $\mathbb{X}$ are not overlapped with those in $\mathbb{X}^{\text{aux}}$.

We argue that the above problem setting is both practically valuable and convenient in implementation. Albeit weakly-supervised learning is preferred for object detection, several well-annotated datasets, *e.g.*, COCO [17], have already been publicly available. Our problem setting allows the learned bounding box adjusters to be deployed in training new classes of object detector, thereby being expected to be advantageous to conventional WSOD solely relying on $\mathbb{X}$. In OCUD [37], the well-annotated dataset $\mathbb{X}^{\text{aux}}$ is directly incorporated with the weakly-annotated dataset $\mathbb{X}$ for WSOD. In our problem setting, the well-annotated dataset $\mathbb{X}^{\text{aux}}$ can be safely abandoned after learning bounding box adjusters. Then, LBBAs can be incorporated with any weakly annotated dataset $\mathbb{X}$ for WSOD. We note that LBBAs can avoid the direct leakage of well-annotated dataset $\mathbb{X}^{\text{aux}}$ to the users with weakly annotated dataset $\mathbb{X}$, thereby being more convenient, economic, and secure in practice.

3.2. Overview

In general, our method involves two subtasks, *i.e.*, (i) learning bounding box adjusters, and (ii) LBBA-boosted WSOD. The overall training procedure is shown in Fig. 1. To better draw the LBBAs from well-annotated auxiliary dataset, we formulate the learning of bounding box adjusters as a bi-level optimization problem. In the lower-subproblem, we use a WSOD method and current LBBA g_t to update the object detection model f_{t+1} from $\{(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}, \mathbf{y}^{\text{aux}})\}$. So the learned f_{t+1} can also be represented as a function of LBBA. Therefore, the upper-subproblem is formulated to learn g_{t+1} specified for optimizing the performance of the weakly-supervised object detector by using the well-annotated data $\{(\mathbf{I}^{\text{aux}}, \{\mathbf{b}^{\text{aux}}\}, \mathbf{y}^{\text{aux}})\}$. In each stage, we first update the learning of bounding box adjuster g_{t+1} by fixing f_t, and then update the weakly-supervised object detector f_{t+1} by fixing LBBA g_{t+1}. With several stages ($T = 3$) of training. We can obtain a set of LBBA models $\{g_0, ..., g_T\}$ with one for each stage.

For LBBA-boosted WSOD, the well-annotated dataset $\mathbb{X}^{\text{aux}}$ can be abandoned, and only the LBBA models $\{g_0, ..., g_T\}$ and the weakly annotated dataset $\mathbb{X}$ are required. LBBA-boosted WSOD also involves several stages (*i.e.*, T). In each stage (*e.g.*, t), we use the current object detector f_t to obtain a set of selected proposals and exploit the stage-wise LBBA g_t for bounding box adjustment. Then, the adjusted bounding boxes are introduced into the WSOD model for updating f_{t+1}. In the following, after introducing the baseline WSOD model used in this work, we present our solutions to the subtasks of both learning bounding box adjusters and LBBA-boosted WSOD in detail.

3.3. Baseline WSOD Model

To learn both bounding box regression and proposal classification from weakly-annotated dataset, we adopt the method proposed in [31, 33] as our baseline network $f(\mathbf{I}, \mathbb{P}; \theta_f)$. Here, θ_f denotes the model parameters of the object detector. Specifically, the network $f(\mathbf{I}, \mathbb{P}; \theta_f)$ involves a basic multi-instance-learning (MIL) branch as well as an independent bounding box regression (BBR) branch. Given an input image $\mathbf{I}$ with image-level label $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_C\}$ as well as R precomputed proposals $\mathbb{P}_{\text{mil}} = \{\mathbf{p}_{\text{mil},1}, ..., \mathbf{p}_{\text{mil},R}\}$, MIL branch generates two $R \times C$ logits $\mathbf{x}^{\text{cls}}$ and $\mathbf{x}^{\text{det}}$, which are passed through softmax layers. Then, a fusion score $\mathbf{s} = \sigma_{\text{cls}}(\mathbf{x}^{\text{cls}}) \cdot \sigma_{\text{det}}(\mathbf{x}^{\text{det}})$ can be computed by performing element-wise product on those of classification and localization. Finally, the image-level score of class c can be attained by

$$\mathbf{q}_c = \sum\nolimits_{i=1}^{R} \mathbf{s}_{i,c}. \tag{1}$$

And the MIL branch can be optimized by

$$\mathcal{L}_{\text{wsddn}} = \text{BCE}(\mathbf{q}, \mathbf{y}; \theta_{\text{f}}), \tag{2}$$

where $\text{BCE}(\cdot,\cdot)$ denotes the binary cross-entropy loss. To improve detection quality, we also introduce pseudo label mining strategy and construct instance refinement branch optimized by a set of weighted instance refinement loss $\mathcal{L}_{\text{r}}$ [26, 25, 20].

In typical single phase WSOD, the precomputed proposals classified to a specific class are taken as the detection results. To improve the object localization performance, we follow [31] to introduce an RPN module into our WSOD network for generating region proposals $\mathbb{P}_{\text{rpn}} = \{\mathbf{p}_{\text{rpn}}\}$. Then, all proposals from $\mathbb{P} = \mathbb{P}_{\text{mil}} \cup \mathbb{P}_{\text{rpn}}$ are sent into bounding box regression branch to generate corresponding localization outputs. Following standard Faster R-CNN [19], both RPN module and bounding box regression branch are trained by the losses $\mathcal{L}_{\text{rpn-cls}}$, $\mathcal{L}_{\text{rpn-det}}$ and $\mathcal{L}_{\text{det}}$ defined on pseudo ground-truth instances selected by refinement scores. Thus, the learning objective of our baseline WSOD model can be written as,

$$\mathcal{L}_{\text{wsod}} = \mathcal{L}_{\text{wsddn}} + \mathcal{L}_{\text{r}} + \mathcal{L}_{\text{rpn-cls}} + \mathcal{L}_{\text{rpn-det}} + \mathcal{L}_{\text{det}}, \tag{3}$$

where $\mathcal{L}_{\text{r}}$ and $\mathcal{L}_{\text{rpn-cls}}$ are the cross-entropy losses supervised by pseudo class labels on the selected proposals, while $\mathcal{L}_{\text{rpn-det}}$ and $\mathcal{L}_{\text{det}}$ are the smooth-L1 losses [7] supervised by the proposal boxes of pseudo ground-truths. Note that we follow the same strategy of OICR [26] to generate pseudo ground-truths.

We note that the bounding box regression branch in baseline WSOD model is learned based on the supervision from the precomputed proposals, which naturally are not precise enough. In the subsequent subsections, we learn a set of bounding box adjusters to provide better ground-truth for supervising the bounding box regression branch, thereby being beneficial to detection performance. Moreover, we use the above baseline WSOD model as an example to show the effectiveness of the learned bounding box adjusters. Actually, our proposed method is independent with most existing WSOD methods and can be incorporated with them to further boost detection performance. And we will illustrate this point in the experiments.

3.4. Learning Bounding Box Adjusters

3.4.1 Bi-level Optimization Formulation

To formulate our weakly supervised object detection problem elegantly, we first revisit the traditional EM algorithm

Algorithm 1 Learning Bounding Box Adjusters

Input: Auxiliary dataset $\mathbb{X}^{\text{aux}}$, adjuster network g , WSOD network f^{aux}, stage num T

Output: Adjuster parameters $\{\theta_g^0 \dots \theta_g^T\}$

Initialize θ_g^0 on $\mathbb{X}^{\text{aux}}$

$\theta_{f^{\text{aux}}}^0 \leftarrow \arg\min_{\theta_{f^{\text{aux}}}} \mathcal{L}_{\text{wsod}} + \mathcal{L}_{\text{bbr}}$

for $t = 0 ... T-1$ **do**

  E-Step:

  $\theta_{\text{g}}^{t+1} \leftarrow \arg\min_{\theta_g} \mathcal{L}_{\text{bba}}$

  M-Step:

  $\theta_{f^{\text{aux}}}^{t+1} \leftarrow \arg\min_{\theta_{f^{\text{aux}}}} \mathcal{L}_{\text{wsod}} + \mathcal{L}_{\text{bbr}}$

return $\{\theta_g^0 \dots \theta_g^T\}$

for weakly supervised learning. In particular, E-step is used to update latent variable $\hat{\text{b}}$,

$$\hat{\text{b}} = \arg\max_{\text{b}_{\text{latent}}} \log P(\mathbf{y}|\text{b}_{\text{latent}}) - \mathcal{L}(\text{b}_{\text{latent}}, f(\mathbf{I}, \mathbb{P}; \theta_f)). \quad (4)$$

For WSOD with box regression, $\mathbf{y}$ is image class labels, $\mathcal{L}$ is defined as box regression loss (*e.g.*, smooth L1 loss [7] for bounding box regression), $\hat{\text{b}}$ means latent bounding box variables, and $P(\mathbf{y}|\text{b}_{\text{latent}})$ is probability of $\mathbf{y}$ with given b_{latent} in WSOD training. And $f(\mathbf{I}, \mathbb{P}; \theta_f)$ is bounding box output from WSOD network f with corresponding parameters θ_f. We mainly discuss $\mathcal{L}$ in next paragraphs. Then, M-step is deployed to update the model parameters θ_f.

$$\theta_f = \arg\min_{\theta_f} \mathcal{L}(\hat{\text{b}}, f(\mathbf{I}, \mathbb{P}; \theta_f)), \quad (5)$$

where $\mathcal{L}$ is a combination of weakly supervised object detection loss $\mathcal{L}_{\text{wsod}}$ and bounding box regression loss $\mathcal{L}_{\text{bbr}}$.

As mentioned above, previous methods utilize precomputed proposals as well as pseudo ground-truth mining in E-step, and then update box regression branch of WSOD network in M-step. However, optimizing $P(\mathbf{y}|\text{b}_{\text{latent}})$ in E-step with only image-level supervision to improve quality of $\hat{\text{b}}$ is difficult. Besides, when optimizing $\mathcal{L}$ in E-step, precomputed proposals are designed for generating region proposals for box regression of object detection, which are not suitable for final object localization. To tackle this problem, we want to use extra well-annotated data to supervise a learnable model, make it generate more precise $\hat{\text{b}}$ in E-step. Therefore, we first introduce a full-annotated auxiliary dataset $\mathbb{X}^{\text{aux}}$ to provide class-agnostic localization supervision. And then, we aim to introduce a class-agnostic Learnable Bounding Box Adjuster (LBBA) $g(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_g)$ trained on $\mathbb{X}^{\text{aux}}$, which takes the selected proposals from $\mathbb{P}^{\text{aux}} = \mathbb{P}^{\text{aux}}_{\text{mil}} \cup \mathbb{P}^{\text{aux}}_{\text{rpn}}$ as the input. For each $\mathbf{p}^{\text{aux}} \in \mathbb{P}^{\text{aux}}$, $g(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_g)$ aims to predict a more precise estimation of bounding box $\hat{\mathbf{b}}^{\text{aux}}$, which is then used to supervise the bounding box regression branch in WSOD. Denoted by $\tilde{\mathbf{b}}^{\text{aux}}$ the output of bounding box regression. We apply smooth L1 loss [7] $\mathcal{L}_{\text{bbr}}$ for supervising bounding box regression branch of f,

$$\mathcal{L}_{\text{bbr}} = \sum\nolimits_{\mathbf{p}^{\text{aux}} \in \mathbb{P}^{\text{aux}}} \text{Smooth}_{L1}(\hat{\mathbf{b}}^{\text{aux}}, \tilde{\mathbf{b}}^{\text{aux}}; \theta_f). \quad (6)$$

Using the ground-truth bounding box $\mathbf{b}^{\text{aux}}$ from $\mathbb{X}^{\text{aux}}$, we further introduce a loss $\mathcal{L}_{\text{bba}}$ for supervising the learning of bounding box adjusters,

$$\mathcal{L}_{\text{bba}} = \sum\nolimits_{\mathbf{p}^{\text{aux}} \in \mathbb{P}^{\text{aux}}} \text{Smooth}_{L1}(\mathbf{b}^{\text{aux}}, \tilde{\mathbf{b}}^{\text{aux}}; \theta_g). \quad (7)$$

To this end, we suggest to utilize LBBA g to generate latent variable $\hat{\text{b}}_{\text{aux}}$ on $\mathbb{X}^{\text{aux}}$.

$$\begin{aligned} \hat{\text{b}}_{\text{aux}} &= g(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_g) \\ \theta_g &= \arg\min_{\theta_g} \mathcal{L}_{\text{bba}}(\{\text{b}^{\text{aux}}\}, g(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_g)) \end{aligned} \quad (8)$$

After introducing LBBA g into WSOD, our WSOD problem can be transferred into a **bi-level optimization problem**, here we state how to build bi-level optimization.

Lower subproblem. During M-step, WSOD network f is supervised by both image class label $\mathbf{y}$ as well as latent variable $\hat{\text{b}}^{\text{aux}}$, which is output of LBBA network $g(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_g)$. Therefore we update parameters of WSOD network $\theta_{f^{\text{aux}}}$ by minimizing $\mathcal{L}_{\text{wsod}} + \mathcal{L}_{\text{bbr}}$, which is shown as follows,

$$\theta_{f^{\text{aux}}} = \arg\min_{\theta_{f^{\text{aux}}}} (\mathcal{L}_{\text{wsod}} + \mathcal{L}_{\text{bbr}})(\hat{\text{b}}^{\text{aux}}, f^{\text{aux}}(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_{f^{\text{aux}}})) \quad (9)$$

Upper subproblem. Taking above equations into consideration, WSOD parameters $\theta_{f^{\text{aux}}}$ can be seen as a function of LBBA parameters θ_g (*i.e.*, $\theta_{f^{\text{aux}}}(\theta_g)$). Thus, in E-step the upper subproblem on θ_g is defined for optimizing $\mathcal{L}_{\text{bba}}$ on the WSOD network $f^{\text{aux}}(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_{f^{\text{aux}}}(\theta_g))$,

$$\theta_g = \arg\min_{\theta_g} \mathcal{L}_{\text{bba}}(\{\text{b}^{\text{aux}}\}, f^{\text{aux}}(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_{f^{\text{aux}}}(\theta_g))) \quad (10)$$

where g generates adjusted bounding box regression for given proposals from WSOD f^{aux}. Thus upper subproblem has transferred into a fully-supervised setting.

3.4.2 EM-like Multi-stage Training Algorithm

From Eqns. (9,10), the direct optimization of θ_g involves the cumbersome computation of the partial gradient $(\partial \mathcal{L}_{\text{bbr}} / \partial \theta_f)(\partial \theta_f / \theta_g)$. Briefly, direct joint training of two networks to solve this bi-level optimization problem is harmful to the generalization ability of LBBA. And EM-like training strategy can keep that of LBBA. Therefore, to avoid this issue, we suggest an EM-like multi-stage training algorithm. Suppose that $f_t(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}_{\text{mil}}; \theta_f^t)$ and $g_t(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_g^t)$ are the learned models at stage t. In the E-step, we use $f_t(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}_{\text{mil}}; \theta_f^t)$ to generate and select the proposals $\mathbb{P}^{\text{aux}}$, which are then deployed to learn $g_{t+1}(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}; \theta_g^{t+1})$. In the M-step, we use θ_g^{t+1} to substitute θ_g in $\mathcal{L}_{\text{bbr}}$, and obtain $f_{t+1}(\mathbf{I}^{\text{aux}}, \mathbb{P}^{\text{aux}}_{\text{mil}}; \theta_f^{t+1})$ by solving the lower subproblem, thereby resulting in our EM-like multi-stage training algorithm. In the following, we explain the initialization, E-step, and M-step in more detail.

Initialization. To begin with, we utilize $\mathbb{X}^{\text{aux}}$ to train a two-stage detector with class-agnostic bounding box regression branch, which is then used as the bounding box adjuster g_0 at stage $t = 0$. Then, the selected proposals from $\mathbb{P}^{\text{aux}}_{\text{mil}}$ are fed into g_0 to generate the adjusted bounding boxes for supervising the learning of WSOD model f_0.

E-step. Given the learned model parameters θ^t_f of f_t at stage t, the E-step aims at learning the bounding box adjuster g_{t+1} with the model parameters θ^{t+1}_g. For an image $\mathbf{I}^{\text{aux}}$ from $\mathbb{X}^{\text{aux}}$, we utilize the RPN module of f_t to generate a set of region proposals $\mathbb{P}^{\text{aux}}_{\text{rpn}}$. We empirically find that it is better to take the region proposal instead of the bounding box predicted by f_t as the input to g_{t+1}. Moreover, both the precomputed and the generated proposals $\mathbb{P}^{\text{aux}}_{\text{mil}} \cup \mathbb{P}^{\text{aux}}_{\text{rpn}}$ are beneficial to the training of g_{t+1}. Thus, we use f_t with the parameters θ^t_f to predict the bounding boxes, and decode them to generate the corresponding selected proposals $\mathbb{P}^{\text{aux}}_{\text{wsod}}$ from $\mathbb{P}^{\text{aux}}_{\text{mil}} \cup \mathbb{P}^{\text{aux}}_{\text{rpn}}$. The model g_{t+1} takes $\mathbb{P}^{\text{aux}}_{\text{wsod}}$ as the input to predict a set of adjusted bounding boxes $\{\hat{\mathbf{b}}^{\text{aux}}\}$. With the ground-truth bounding boxes from $\mathbb{X}^{\text{aux}}$, we train the bounding box adjuster g_{t+1} with the parameters θ^{t+1}_g at stage $t+1$ by minimizing the loss $\mathcal{L}_{\text{bba}}$.

M-step. With the help of the learned model parameters θ^{t+1}_g of g_{t+1}, the M-step learns the WSOD model f_{t+1} with the model parameters θ^{t+1}_f. In the forward propagation, an image $\mathbf{I}^{\text{aux}}$ from $\mathbb{X}^{\text{aux}}$ is fed into the current WSOD model to generate a number of region proposals $\mathbb{P}^{\text{aux}}_{\text{rpn}}$ and bounding boxes. Then, we decode the predicted bounding boxes to obtain the selected proposals $\mathbb{P}^{\text{aux}}_{\text{wsod}}$ from $\mathbb{P}^{\text{aux}}_{\text{mil}} \cup \mathbb{P}^{\text{aux}}_{\text{rpn}}$. Taking $\mathbb{P}^{\text{aux}}_{\text{wsod}}$ as the input, the adjusted bounding boxes predicted by the LBBA g_{t+1} are then used to define the loss $\mathcal{L}_{\text{bbr}}$. Finally, the WSOD model f_{t+1} with the model parameters θ^{t+1}_f can be trained by minimizing the combined loss $\mathcal{L}_{\text{wsod}} + \mathcal{L}_{\text{bbr}}$.

To sum up, after the initialization, our training algorithm alternates between the E-step and M-step for T times. Hence, it is a multi-stage training scheme, where we run the E-step and M-step once in each stage. The training process of LBBA is given in Algorithm 1.

3.5. LBBA-boosted WSOD

After learning bounding box adjusters, the well-annotated auxiliary dataset can be abandoned. For the LBBA-boosted WSOD task, we only require a weakly-annotated dataset $\mathbb{X}$ as well as a set of learned bounding box adjusters $\{g_0, ..., g_T\}$. The multi-stage scheme is also adopted to train WSOD, and we use stage t as an example to illustrate the training process. In particular, an image $\mathbf{I}$ from $\mathbb{X}$ is fed into the current WSOD model to generate a number of region proposals $\mathbb{P}_{\text{rpn}}$ and bounding boxes. Then, we decode the predicted bounding boxes to obtain the selected proposals $\mathbb{P}_{\text{wsod}}$ from $\mathbb{P}_{\text{mil}} \cup \mathbb{P}_{\text{rpn}}$. Taking $\mathbb{P}_{\text{wsod}}$ as the input, the adjusted bounding boxes predicted by the LBBA g_t are then used to define the loss $\mathcal{L}_{\text{bbr}}$. Finally, the WSOD model f_t with the model parameters θ^t_f can be trained by minimizing the combined loss $\mathcal{L}_{\text{wsod}} + \mathcal{L}_{\text{bbr}}$. After T stages of training, the WSOD model at stage T, *i.e.*, f_T with parameters θ^T_f, can be kept and applied to the test images. The training process of LBBA-boosted WSOD is given in Algorithm 2.

Nonetheless, we empirically find that updating WSOD network with only the last g_T can attain a similar performance. Hence we can build a lighter pipeline by only using the last g_T.

Algorithm 2 LBBA-boosted WSOD

Input: Weakly-annotated dataset $\mathbb{X}$, stage num T, adjuster network g, adjuster parameters $\{\theta^0_g \dots \theta^T_g\}$, WSOD network f

Output: WSOD network parameters θ^T_f

1: **for** $t = 0...T$ **do**
2: $\quad \theta_g \leftarrow \theta^t_g$
3: $\quad \theta^t_f \leftarrow \arg\min_{\theta_f} \mathcal{L}_{\text{wsod}} + \mathcal{L}_{\text{bbr}}$
4: **return** θ^T_f

3.6. Masking Strategy for Proposal Classification

The above training algorithm can improve localization ability of WSOD network but cannot improve the ability of proposal classification. To further improve the detection performance, we introduce an additional multi-label image classifier $h(\mathbf{I}; \theta_h)$ and present a classification score masking strategy. During training, we utilize images and corresponding image labels of dataset $\mathbb{X}$ to train h; during testing, given input image $\mathbf{I}$, we obtain image classification score by $\hat{\mathbf{s}} = h(\mathbf{I}; \theta_h)$, where $\hat{\mathbf{s}} \in \mathbb{R}^{1 \times C}$ is per-class prediction scores of $\mathbf{I}$. Therefore, we can judge which categories should not be included in $\mathbf{I}$, and suppress the corresponding output of WSOD. Specifically, we select a threshold τ (*i.e.*, = -3.0), if $\hat{\text{s}}_c < \tau$, we assert that the category c is not appeared in this image. Therefore, for each category c with $\hat{s}_c < \tau$, score of i-th proposal $\hat{\mathbf{b}}_{i,c}$ is set to 0 to eliminate wrong predictions.

4. Experiments

4.1. Datasets and Evaluation Metrics

Auxiliary Dataset. MS-COCO 2017 [17] is a large-scale object detection dataset. Note that MS-COCO dataset includes 80 different object classes. To eliminate semantic overlap and show the generalization ability of our method, we construct a subset of MS-COCO by excluding PASCAL VOC classes instance annotations and call it COCO-60. As such, COCO-60 dataset contains $\sim$98K training images and $\sim$4K validation images, respectively.

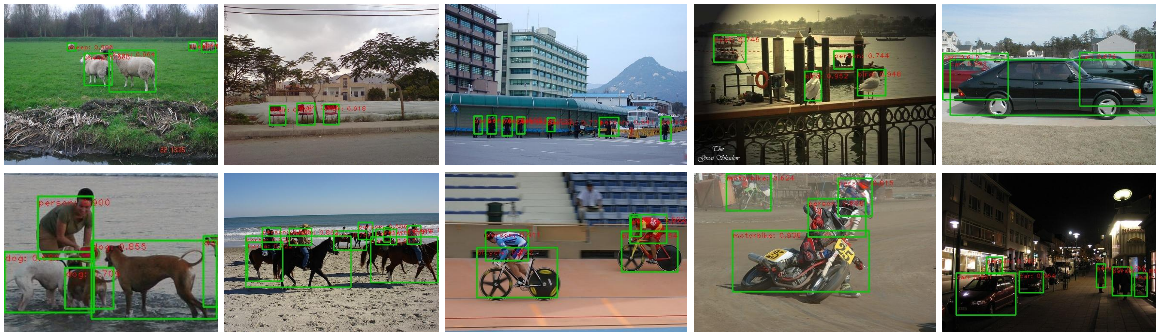

Figure 2. Visualization results of our method on PASCAL VOC 2007, which has the ability to generate precise bounding boxes.

Table 1. Single model detection results on PASCAL VOC 2007 and 2012, where $^+$ means the results with multi-scale testing, * means joint training of WSOD models on auxiliary dataset and weakly-annotated dataset.

Methods	mAP (07)	mAP (12)
OICR$^+$ [26]	41.2	37.9
PCL$^+$ [25]	43.5	40.6
Yang *et al.*$^+$ [33]	51.5	46.8
WSOD 2$^+$ [34]	53.6	47.2
Arun *et al.* [2]	52.9	48.4
C-MIDN$^+$ [32]	52.6	50.2
MIST (Full)$^+$ [20]	54.9	52.1
MSD-Ens$^+$ [14]	51.1	-
OICR+UBBR [12]	52.0	-
Zhong *et al.* (R50-C4)* [37]	**57.8**	-
Zhong *et al.* (R50-C4)$^{+*}$ [37]	**59.7**	-
Ours	**56.5**	**54.7**
Ours$^+$	**56.6**	**55.4**
Upper bounds:		
Faster R-CNN [19]	69.9	67.0

Table 2. Single model correct localization (CorLoc) results on PASCAL VOC 2007 and 2012, where $^+$ means the results with multi-scale testing, * means joint training of WSOD models on auxiliary dataset and weakly-annotated dataset.

Methods	CorLoc (07)	CorLoc (12)
OICR$^+$ [26]	60.6	62.1
PCL$^+$ [25]	62.7	63.2
Li$^+$ [13]	68.6	67.9
Yang *et al.*$^+$ [33]	68.0	69.5
WSOD 2$^+$ [34]	69.5	71.9
Arun *et al.*[2]	70.9	69.5
C-MIL$^+$ [30]	65.0	67.4
MIST (Full)$^+$ [20]	68.8	70.9
WSLAT-Ens [21]	58.8	-
MSD-Ens$^+$ [14]	66.8	-
OICR+UBBR [12]	47.6	-
Zhong *et al.* (R50-C4)* [37]	**73.6**	-
Zhong *et al.* (R50-C4)$^{+*}$ [37]	**74.4**	-
Ours	**72.3**	**73.2**
Ours$^+$	**72.5**	**73.7**

Target Datasets. PASCAL VOC 2007 and 2012 datasets contain 9,963 images and 22,531 images collected from 20 object classes. For fair comparison, we use *trainval* set for training WSOD networks and report evaluation results on *test* set. During the training process, only image-level labels are used as supervision. We also utilized other datasets to evaluate our LBBA, see the suppl. for details.

Evaluation Metrics. Since our method aims at improving object detection performance, Average Precision (AP) is used as the basic evaluation metric in our experiments. We also adopt CorLoc [4] as another evaluation metric.

4.2. Comparison with State-of-the-arts

We *state the implementation details in the suppl.* and we build up all experiments based on it. We compare our method with several state-of-the-art WSOD approaches in terms of detection and localization performance on PASCAL VOC datasets. As suggested in [3, 26, 25, 33, 20, 2, 37], we report detection results on *test* set and localization results on *trainval* set, respectively. Table 1 compares the results of different state-of-the-art WSOD approaches on PASCAL VOC 2007 and 2012 datasets. It can be seen that our LBBA improves OICR and OICR+REG over 15.3% and 5.0% on PASCAL VOC 2007 dataset, respectively. Furthermore, our method performs better than all competing methods, except Zhong *et al.* [37]. Note that [37] uses stronger backbone model and knowledge transfer strategy by directly incorporating source and target datasets. Moreover, the auxiliary dataset adopted in Zhong *et al.* is different from ours (See the suppl. for more details). As shown in Fig. 2, our method has the ability to generate precise bounding boxes. On PASCAL VOC 2012, our LBBA is superior to all competing methods and obtains more than 1% gains over all WSOD approaches. Experimental results show that our method is effective in improving the detection performance of WSOD.

We further evaluate the localization performance of our method. Table 2 lists results of several state-of-the-art WSOD approaches on PASCAL VOC 2007 and 2012. Our LBBA outperforms OICR by 11.7% and also improves the baseline OICR+REG over 4.3% on PASCAL VOC 2007 dataset. Besides, our LBBA performs better than all competing methods. Meanwhile, on PASCAL VOC 2012, our LBBA is also superior to all competing methods, and obtains 1.3% gain over WSOD 2[34]. In comparison to Zhong *et al.* [37], our LBBA-based method employs a weaker backbone model and avoids the direct joint use of the source and target datasets, while still achieving competitive Cor-

Table 3. Comparison of different backbone models of Adjuster g on VOC 07, where iterations T of multi-stage learning is set to 3 while WSDDN [3] is used as WSOD network f.

Backbone of Adjuster g	mAP (VOC 07)	CorLoc (VOC 07)
VGG-16	50.2	67.7
R50-C4	**52.7**	**70.3**

Table 4. Comparison of various WSOD networks f on VOC 07.

Method	mAP (VOC 07)	CorLoc (VOC 07)
Baseline (WSDDN)	46.6	64.7
Baseline (OICR)	48.6	66.8
Baseline (OICR+[20])	51.4	64.9
Ours (WSDDN)	**52.7**	**70.3**
Ours (OICR)	**55.1**	**71.0**
Ours (OICR+[20])	**55.8**	**71.6**

Loc results under the settings of both single-scale testing and multi-scale testing. The above results show that our LBBA-based method is effective in improving the localization performance of WSOD.

4.3. Ablation Study

Additionally, we employ PASCAL VOC 2007 to assess the effect of some key components on our LBBA. We state a more detailed ablation study in the suppl..

Backbone Models of Adjuster g. In this work, Faster R-CNN [19] is used as adjuster. Here, we first evaluate the effect of backbone models on adjuster g. To this end, we compare two CNN architectures as backbone models of Faster R-CNN, i.e., ResNet-50 [10] and VGG-16 [23]. Particularly, we set iterations T of multi-stage learning to 3 and adopt WSDDN [3] as WSOD network f. The compared results on VOC 07 are listed in Table 3, from which we can see that adjuster g with backbone of ResNet-50 outperforms one with backbone of VGG-16 by 2.5% and 2.6% in terms of mAP and CorLoc, respectively. These results show that our method can benefit from a stronger adjuster, which encourages us to develop more effective adjusters.

Effect of WSOD network f. After determining backbone model of adjuster g, we access the impact of WSOD network f. Specifically, we consider three methods (i.e., WSDDN+REG [3], OICR+REG [26] and OICR+REG with top p% pseudo label mining [20]) for our WSOD network f, and compare our LBBA with the original methods (i.e., baseline). The iterations T of multi-stage learning is set to 3, and the results of different WSOD networks f are given in Table 4. First, our LBBA achieves clear performance gains (more than 4%) over the baseline methods for all choices of WSOD networks in terms of mAP and CorLoc. It demonstrates that the proposed LBBA methods can be well generalized to various WSOD networks. Second, our LBBA benefits from stronger WSOD networks, and so we compare with state-of-the-arts by using OICR+[20] as WSOD network f.

Multi-stage LBBAs. The proposed multi-stage learning strategy of LBBAs involves two core factors, *i.e.*, number of iterations (T) and learnable, auxiliary WSOD network f^{aux}. By fixing WSOD network f and adjuster g respectively be OICR+[20] and Faster R-CNN with backbone of ResNet-50, we assess the effects of number of iterations (T) and f^{aux} on our LBBA method. To this end, we learn bounding box adjusters by setting T from 0 to 3. Besides, we replace learnable f^{aux} by using MCG to generate proposals, namely LBBA-MCG. Table 5 gives the results of adjuster g and WSOD network f on COCO-60 and VOC 07 using different learning strategies, respectively. It can be seen that increasing iterations (T) can improve performance of both adjuster g and WSOD network f. However, performance of WSOD network f is sightly improved, when number of iterations $T > 2$. Therefore, $T = 3$ is a good choice to balance efficiency and effectiveness. These results clearly demonstrate the effectiveness of our multi-stage learning strategy. The learnable f^{aux} with 3 iterations is superior to LBBA-MCG by 1.3% and 1.5% for adjuster g and WSOD network f, showing the significance of learnable f^{aux}.

Table 5. Results of adjuster g and WSOD network f on COCO-60 and VOC 07 using different learning strategies, respectively

Learning Strategy	Adjuster mAP (COCO-60)	mAP (VOC 07)
T=0	29.1	53.1
T=1	29.6	54.9
T=2	29.9	55.7
T=3	**30.9**	**55.8**
LBBA-MCG	29.6	54.3

5. Conclusion

In this paper, we presented a knowledge transfer based WSOD method. Our proposed method involves two subtasks, *i.e.*, learning bounding box adjusters and LBBA-boosted WSOD. For the former subtask, we suggested a bi-level optimization formulation on the auxiliary dataset and an EM-like training algorithm to learn multi-stage and class-agnostic LBBAs specified for optimizing WSOD performance. For the later subtask, we adopted a multi-stage scheme to utilize only the LBBAs and weakly-annotated dataset for WSOD. Additionally, a masking strategy is adopted to improve proposal classification for benefiting detection performance. Experimental results show that our proposed method performs favorably against the state-of-the-art WSOD methods and knowledge transfer model with similar problem setting [12, 14, 21, 37]. Nonetheless, we mainly focus on transferring across classes in this paper, while the transferring across domains is not specifically considered. In the future, we will explore suitable domain generalization methods for coping with this issue.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under grant No.s U19A2073, 62006064 and 61806140, and Natural Science Foundation of Tianjin under grant No. 20JCQNJC1530.

References

[1] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *Computer Vision and Pattern Recognition*, 2014.

[2] Aditya Arun, C.V. Jawahar, and M. Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[4] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100, 12 2012.

[5] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89:31–71, 1997.

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[7] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] M. Guillaumin and V. Ferrari. Large-scale knowledge transfer for object localization in imagenet. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3202–3209, 2012.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544, 2014.

[12] Seungkwan Lee, Suha Kwak, and Minsu Cho. Universal bounding box regression and its applications. In C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 373–387, Cham, 2019. Springer International Publishing.

[13] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[14] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Mixed supervised object detection with robust objectness transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 02 2018.

[15] Chenhao Lin, Siwen Wang, Dongqi Xu, Yu Lu, and Wayne Zhang. Object instance mining for weakly supervised object detection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11482–11489. AAAI Press, 2020.

[16] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[20] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] M. Rochan and Y. Wang. Weakly supervised localization of novel objects using appearance transfer. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4315–4324, 2015.

[22] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[24] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[25] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):176–191, 2018.

[26] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2017.

[27] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2016.

[28] Jasper Uijlings, Stefan Popov, and Vittorio Ferrari. Revisiting knowledge transfer for training object class detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1101–1110, 2018.

[29] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[30] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[31] Jiajie Wang, Jiangchao Yao, Ya Zhang, and Rui Zhang. Collaborative learning for weakly supervised object detection. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 971–977. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[32] G. Yan, B. Liu, N. Guo, X. Ye, F. Wan, H. You, and D. Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9833–9842, 2019.

[33] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8372–8381, 2019.

[34] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[35] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Q. Tian. Zigzag learning for weakly supervised object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4262–4270, 2018.

[36] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 928–936, 2018.

[37] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 615–631, Cham, 2020. Springer International Publishing.

[38] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.