Boosting Weakly-Supervised Object Detection via Online Nesting Boxes Regression

Anonymous CVPR submission

Paper ID 3589

Abstract

Weakly-supervised object detection requires localizing multiple object instances (e.g., birds or dogs) in an image with only image-level supervision. Existing approaches adopt the paradigm of multi-instance learning and tend to select the most discriminative bounding box (by classification score) as an object instance from thousands of region proposals. However, such approaches neglect learning regression capability, which plays a key role in refining detection results. In this paper, we propose a novel online nesting boxes regression (ONBR) network for weakly-supervised object detection. In particular, we first infer highly-confident boxes from weak supervision by multi-instance learning. For each box, we further seek inner and outer boxes by following a rule of nesting boxes mining. Second, we consider these mined nesting boxes as pseudo ground truths, and jointly refine region proposals by instance-level classification and regression. Third, once the instance classifier is updated, new nesting boxes can be discovered in an iterative way. An empirical study indicates that the proposed ONBR-Network can suppress partial and oversized detection results of an object instance. We further evaluate our approach on two object detection benchmarks, i.e., PASCAL VOC 2007/2012, respectively. Experiment results show that the proposed approach can boost 5.5% *mAP by a single model on PASCAL VOC 2007, compared with the state-of-the-art results under the same setting.*

1. Introduction

The capability of recognizing and localizing objects in an image has drawn great attention in recent years. Significant progresses have been achieved with the development of convolutional neural network [4, 14, 11, 22]. However, current state-of-the-art object detectors usually require large-scale training datasets with bounding box annotation (e.g. PASCAL VOC [7], MS COCO [17], Open Image[15]). To relieve heavy manual labeling effort, weakly-supervised object detection paradigm has been proposed by using only image-level annotation.

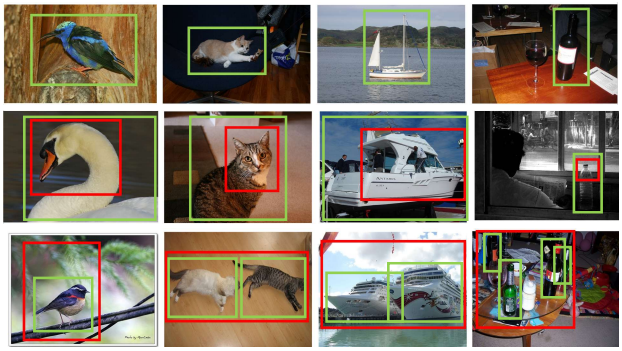

Figure 1. Typical weakly-supervised object detection results. The green boxes in the first row indicate correct detection. The red boxes in the second/third row indicate partial/oversized results, respectively. By seeking inner or outer boxes by the proposed ONBR-network, more correct detection boxes can be discovered.

The challenges of weakly-supervised object detection (WSOD) mainly come from two folds. First, most previous works adopt multiple instance learning method to transform WSOD into multi-label classification problem. However, a classification model often targets at judging the existence of objects for a category, while is not able to predict the location, size and the number of objects in images. Therefore, there are usually partial or oversized detection results produced by WSOD approaches, as shown in Figure 1. Second, bounding box regression has been considered as an important component in typical fully-supervised object detection (FSOD), as it is able to reduce the localization errors by refining region proposals. However, it is difficult to integrate such a key module into WSOD framework, because no additional bounding box annotation can be utilized as regression supervision. Therefore, there is still large performance gap between WSOD and FSOD.

To relieve the above challenges, recent progress has been made into two dimensions. First, existing works have pro-

posed different types of instance refinement algorithms, for better discovering correct object instances. OICR determines boxes with the highest confidence as positive samples after several rounds of refinement [2, 23]. W2F proposes a bounding box selection strategy based on the prediction results of OICR, and further selects positive samples for the second fully-supervised detector training by thresholding, NMS and merging [29]. Second, some works have tried to improve the regression ability of WSOD. Most of these works propose to train a weakly-supervised detector in an initial stage, and further use the produced outputs as pseudo ground truths to train another fully supervised detector [8, 19, 6, 27]. Boundary regression is applied in the second detector. However, the performance of the above two refinement-based approaches heavily relies on the accuracy of the initial object detection results, which limits further improvement with large margins.

In this paper, we propose a novel weakly-supervised object detection approach by online nesting boxes regression (ONBR). First, given an input image with thousands of region proposals (e.g., generated by selective search [26]), we follow existing works of WSOD to learn an initial instance classifier by multi-instance learning [18, 1]. Such a classifier can help to select multiple highly-confident bounding boxes as seeds (i.e., possible object instances) by adopting threshold and non-maximum suppression. Second, as partial or oversized object detection results are usually produced by WSOD approaches, we further seek inner and outer boxes by following a rule of nesting boxes mining (NBM). The goal of NBM is to discover spatially-nesting regions with adequate discrimination ability, although the scores of these regions (generated by the above instance classifier) are not the highest. Such a design helps to prevent WSOD from overfitting initial seeds, and to cover complete object instances as much as possible. Third, we consider these mined nesting boxes as pseudo ground truths, and progressively refine the instance classifier, and thus more accurate seeds with their nesting boxes can be generated in return. An empirical study indicates that this refinement progress can gradually suppress partial and oversized region proposals, and thus one bounding box (which is close to a complete object instance) from a nesting boxes structure can be finally discovered. Finally, we integrate a bounding box regressor into WSOD to predict coordinates offsets and fine-tune the location of each proposal, by taking the nesting boxes as regression supervision. Such a design enables each proposal to regress to its pseudo ground truths nearby by calculating smooth L_1 loss.

To the best of our knowledge, this work is the first to explore the possibility of learning regression ability in weakly-supervised object detection. The contributions can be summarized as follows:

1. We propose to integrate bounding box regression and progressive instance classifier training strategies into WSOD framework in an end-to-end fashion.
2. We propose a nesting boxes mining strategy to suppress partial and oversized region proposals, which tends to regress bounding boxes to complete objects, even without bounding box annotations.
3. We achieve 53.2% and 48.6% mAP on PASCAL VOC 2007 and 2012 benchmarks respectively, which shows the effectiveness of the proposed ONBR-Network.

2. Related Works

2.1. Fully-supervised Object Detection

Various CNN based object detectors are proposed in recent years with the development of deep learning, and the success of convolutional neural network in classification. R-CNN innovatively proposes to use proposal-based methods which has pushed detection results on PASCAL VOC dataset to a new level [9]. Fast R-CNN proposes RoI Pooling to significantly reduce the redundant convolutional calculation in R-CNN [8]. Region proposal network (RPN) is proposed to combine the proposal stage and classification stage together, and make the whole network trainable in an end-to-end way [19]. A number of promising weakly-supervised detectors are proposed by inspiring from the above two-stage detectors, and adopt similar architectures in their framework. In particular, proposals are first generated by rule-based methods like selective search [26] and EdgeBoxes [32]. Region features can be further extracted by ROI Pooling [8] or SPP [10] layer. The difference among different approaches mainly comes from the design of head networks and loss functions.

2.2. Weakly-supervised Object Detection

Weakly-supervised object detection has attracted great attention in recent years, as it has no need for expensive and time-consuming human labeling efforts. Existing works on WSOD can be roughly grouped into multiple-instance learning (MIL) based methods and progressive-learning based methods. On the one hand, MIL algorithms usually transform WSOD into a multi-label classification problem [25, 12, 16, 21]. Hakan et al. proposes WSDDN which performs multiplication on the score of a classification and a detection branch, so that high-confidence positive samples can be selected [3]. Peng et al. proposes OICR which is able to online refine instance classifiers based on the output of WSDDN, and thus OICR can achieve higher precision. On the other hand, progressive-learning based methods propose to train a fully-supervised object detector by using the outputs generated by MIL-based methods as pseudo ground truth [3]. As class activation map (CAM) [30] produced by a classifier can roughly localize the object, some works try to utilize the CAM to generate coarse detection results

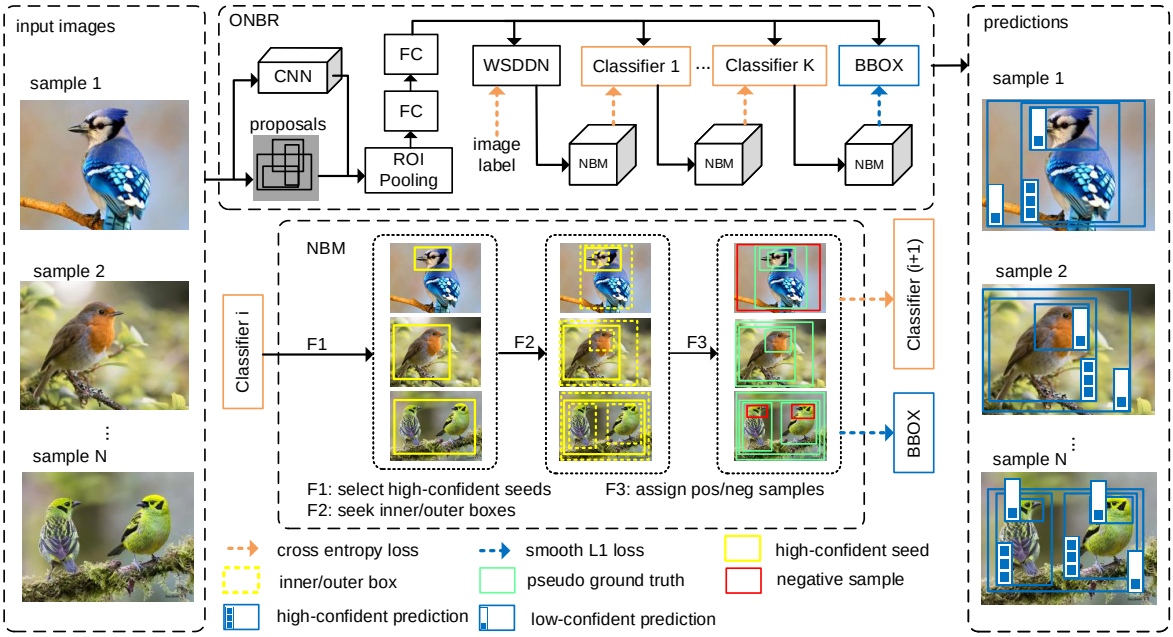

Figure 2. The framework of the proposed online nesting boxes regression (ONBR) network for weakly-supervised object detection. Taking an image as input, ONBR-network consists of a region proposal network, a series of progressively-learned instance classifiers which are updated by the proposed nesting boxes mining (NBM) module, and the final bounding box regression which can fine-tune predicted bounding boxes to target pseudo ground truths. The regression capability poses a significant difference between the proposed ONBR-network and most existing WSOD approaches, in which bounding boxes with confidence scores can be generated as shown in the blue boxes in the right predictions module. The NBM module is designed to discover both pseudo ground truth bounding boxes (in green), as well as negative samples (in red). An object instance can be probably selected as ground truth by the effective nesting boxes strategy, which can extend a seed detection result (solid yellow boxes) with inner/outer boxes (dotted yellow boxes). [best view in color]

[31], or use it as reference for the later refinement operation [28]. Some works find that using the outputs of weakly-supervised object detectors as pseudo ground truths to train another fully-supervised detector, like Fast R-CNN, can always increase the detection performance [2]. Yongqiang et al. focus on the selection strategy of pseudo ground truths, and proposes W2F in to generate more accurate ones so that the training of second detector can benefit from it [29]. The above works have been putting much efforts to mine object instances. However, how to distinguish an complete object instance with its partial or oversized detection results is still a challenge problem. In this paper, we propose an online nesting boxes regression network to mine positive samples, and integrate a boundary regressor into object detectors to further boost the performance.

3. Approach

As illustrated in Figure 2, given an input image, we first feed it into a CNN backbone, and extract its region proposals by selective search [26]. The region features are warped into fixed size by RoI pooling layer, and further embedded by two fully-connected layers. In this section, we introduce the multiple-instance detector which is trained by multiple-instance learning in Section 3.1. We explain the detail of nesting boxes mining which is used for positive samples selection in Section 3.2. We show how to integrate bounding box regressor into weakly-supervised detector in Section 3.3. And finally, we will give an empirical study to analyse the optimization progress of ONBR in Section 3.4.

3.1. Multiple-instance Detector

In weakly-supervised detection, only image-level annotations are available, hence no information can be utilized to apply instance-level supervision in the training stage. We follow WSDDN [3] to transform the WSOD into a mulit-label classification problem, and initially obtain the coarse detection result.

WSDDN is a kind of multiple-instance learning algorithm. We consider each image as a set of bags, and objects as instances. A bag will be labeled positive if at least one instance is positive, and negative otherwise. Given an image, we denote the set of its region proposals as R. The classification and detection features of all regions can be represented as two matrices $\boldsymbol{x}^c, \boldsymbol{x}^d \in \mathbb{R}^{C\times|R|}$ respectively, where C denotes the class number and $|R|$ denotes the number of region proposals. Latter, $\boldsymbol{x}^c$ and $\boldsymbol{x}^d$ will be passed through two softmax operator along two different directions, respec-

Algorithm 1 Nesting boxes mining

Input: All region proposals R, seed region proposals R_{pos}, IoU threshold T_{iou}
Output: R_{pos}

$R_{in} \Leftarrow \emptyset, R_{out} \Leftarrow \emptyset$
for $i = 1$ to $|R_{pos}|$ **do**
  $out\_flag \Leftarrow False$
  for $j = 1$ to $|R|$ **do**
    if $IoU(R_{pos}[i], R[j]) < T_{iou}$ **then**
      if $\exists_R \in R$ satisfy $(IoU(R, R_{pos}[i]) > T_{iou}$ and $IoU(R, R[j]) > T_{iou})$ **then**
        continue
      if $R[j]$ contains $R_{pos}[i]$ **then**
        if $\forall_R \in R_{in}$ satisfy $IoU(R, R[j]) < T_{iou}$ **then**
          $R_{in} = R_{in} \cup \{R[j]\}$
      else if $R_{pos}[i]$ contains $R[j]$ **then**
        if $out\_flag = False$ **then**
          $out\_flag \Leftarrow True$
          $R_{out} \Leftarrow R_{out} \cup \{R[j]\}$
$R_{pos} \Leftarrow R_{pos} \cup R_{in} \cup R_{out}$
return R_{pos}

tively, denoted as

$$[\sigma_{cls}(\boldsymbol{x}^c)]_{ij} = \frac{e^{\boldsymbol{x}^c_{ij}}}{\sum_{k=1}^{C} e^{\boldsymbol{x}^c_{kj}}}, [\sigma_{det}(\boldsymbol{x}^d)]_{ij} = \frac{e^{\boldsymbol{x}^d_{ij}}}{\sum_{k=1}^{|R|} e^{\boldsymbol{x}^d_{ik}}}. \tag{1}$$

We perform element-wise product operation on the two matrices to achieve region-level scores by

$$\boldsymbol{x}^R = \sigma_{cls}(\boldsymbol{x}^c) \odot \sigma_{det}(\boldsymbol{x}^d), \tag{2}$$

and then reduce $\boldsymbol{x}^R$ into an image-level score vector $\boldsymbol{\phi} = [\phi_1, \phi_2, \cdots, \phi_C]$ by summing up over the region dimension: $\phi_c = \sum_{r=1}^{|R|} x^R_{cr}$.

We apply binary cross entropy loss on $\boldsymbol{\phi}$ to optimize the network. The loss function can be formulated as:

$$L_c = -\sum_{c=1}^{C} \{y_c log(\phi_c) + (1 - y_c) log(1 - \phi_c)\}, \tag{3}$$

where y_c indicates the ground truth label of the c^{th} class, where $y_c = 1$ means the input image contains the c^{th} class, and $y_c = 0$ otherwise.

Inspired by OICR proposed in [2], we extend a series of instance classifiers from the region features for further refinement. All these classifiers have the same formulation, and the outputs of previous classifier will supervise the learning of current one.

For the k^{th} classifier, we embed the region feature into $\{C+1\}$-dimension by a fully connected layer, and denote

Algorithm 2 Pseudo classification ground truths assignment

Input: $R, R_{pos}, \boldsymbol{y} = \{y_1, y_2, \cdots, y_c\}, T_{iou}, \boldsymbol{x}^{(k-1)}$
Output: $\boldsymbol{y^k} = \{y^k_{1r}, y^k_{2r}, \cdots, y^k_{cr}\}$

set all $y^k_{cr} = 0$
select the prediction vectors from $\boldsymbol{x}^{(k-1)}$ of all regions in R_{pos} to construct $\boldsymbol{x}^{(k-1)}_{pos}$
for $c = 1$ to C **do**
  if $y_c = 1$ **then**
    for $j = 1$ to $|R|$ **do**
      $max\_ovlap = -1$
      for $i = 1$ to $|R_{pos}|$ **do**
        $ovlap = IoU(R[j], R_{pos}[i])$
        if $ovlap > max\_ovlap$ **then**
          $max\_ovlap = ovlap$
          $w^k_r = \boldsymbol{x}^{(k-1)}_{pos}[ci]$
          if $ovlap > T_{iou}$ **then**
            $y^k_{cr} = 1$
return $\boldsymbol{y^k}$

the embedded feature matrix as $\boldsymbol{x}^k \in \mathbb{R}^{(C+1)\times|R|}$, where C means the class number, and we add one additional dimension for background class. We define the refinement loss function as:

$$L_{rk} = (-\frac{1}{|R|} \sum_{r=1}^{|R|} \sum_{c=1}^{C+1} w^k_r y^k_{cr} log(\boldsymbol{x}^{nk}_{cr})), \tag{4}$$

where w^k_r and y^k_{cr} indicate the loss weight and class label respectively. The values of these two variables are related to the positive samples selection strategy. OICR considers the highest-confident boxes in images as pseudo ground truth, and apply classification loss on those boxes nearby to them. However, such strategy has some limitations, and we replace it as our proposed nesting boxes mining module. The detail of nesting boxes mining and the calculation of w^k_r and y^k_{cr} will be illustrated in Sec. 3.2.

3.2. Nesting Boxes Mining

Existing works most adopt the highly-confident bounding boxes detected by multiple-instance detector as final predictions, or store the results for further use. However, since the multiple-instance detector is trained with only image-level labels, it is not capable enough of distinguishing among tight, partial and oversized boxes. Some detection cases can be found in Figure. 1. To tackle this issue, we propose a simple yet effective module called nesting boxes mining (NBM) to online select positive samples and assign classification label and regression target for each proposal.

Once we obtain the output of an instance classifier, we can filter up the highly-confident boxes after adopting threshold, sorting and non-maximum suppression. The kept

boxes will be considered as seed positive samples. We denote the set of these boxes as R_{pos}. We extend R_{pos} from R by applying Algorithm. 1.

Specifically, we seek the discriminate inner and outer boxes of each seed box in R_{pos} if the both two following **NBM rules** are satisfied.

1. The IoU with seed box smaller than a pre-defined threshold.
2. It does not exist any box which have large IoU with any two boxes in R_{pos} at same time.

We consider the all boxes in extended R_{pos} as pseudo ground truths, and assign classification label to each proposal. The assignment progress is illustrated in Algorithm. 2. Variable w_r^k and y_{cr}^k will be calculated in to the 11^{th} and 13^th line, respectively.

3.3. Bounding Boxes Regressor

Bounding boxes regressor is a necessary component in typical fully-supervised detector, as it can fine tune the positions and sizes of proposals. However, such module is abandoned by most existing multi instances learning-based weakly-supervised detectors. Therefore, an detector is essentially perform a kind of ranking on region proposals. Such methods rely heavily on the accuracy and recall of the proposals generation algorithm. We integrate bounding boxes regressor into our framework to reduce the localization error.

Our integrated bounding box regressor have the same architecture as in Fast R-CNN [8]. We regress 4 coordinates for each proposal following by:

$$\begin{aligned} t_x = (x - x_p)/w_p, t_y = (y - y_p)/h_p \\ t_w = log(w/w_p), t_h = log(h/h_p) \end{aligned}, \tag{5}$$

where x, y, w, h denote the center coordinates, width and height of predicted box, and similarly x_p, y_p, w_p, h_p represented the position and size of the proposal box.

For the backward progress, the bounding boxes regression loss have the following formulation:

$$L_b = \sum_{r=1}^{|R|} b_r \cdot smooth\_l1(t_r, t_r^*), \tag{6}$$

where t_r denotes the 4 predicted coordinates, and t_r^* indicates its corresponding ground truth. Variable b_r represents the loss weight. We adopt the positive samples selected by last NBM as the regression ground truths to make the bounding box regressor trainable. To analyse the gradient back propagation, we re-formulate L_b into:

$$L_b = \sum_{n=1}^{N} \sum_{r=1}^{|R_n|} b_n \cdot smooth\_l1(t_{nr}, t_n^*), \tag{7}$$

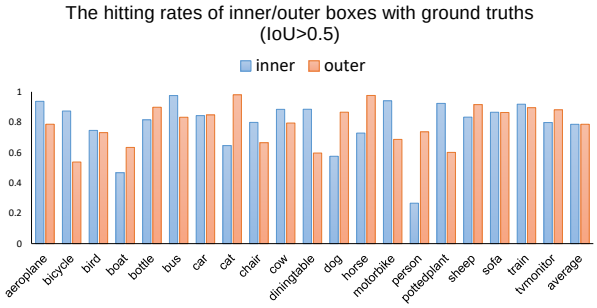

Figure 3. The hitting rate of inner/outer boxes with ground truths on PASCAL VOC 2007 trainval set . The average hitting rates of inner and outer boxes can both achieve 0.78.

where N indicates the number of ground truth instances, and R_n denotes the region proposals which have $IoU > 0.5$ with the n^{th} ground truth box. We set b_n as classification score of the corresponding ground truth box.

3.4. Optimization Analysis

Loss Function. The three above components of online nesting boxes regression (ONBR) network are close related to each other. The outputs of instance classifiers serve as pseudo ground truths and supervise the training of next classifier and bounding box regressor. The whole network can be train in an end-to-end fashion. The overall loss function of ONBR-Net can formulated as:

$$L = L_c + \lambda_1 \sum_{i=1}^{K} L_{rk} + \lambda_2 L_b, \tag{8}$$

where λ_1 and λ_2 are hyparameters, and we adopt $\lambda_1 = 1$ and $\lambda_2 = 0.3$ in our experiments.

Optimization of Multiple-instance Detector. At the beginning of the training stage, most of the region proposals will be classified into background category, thus their foreground probabilities are relatively low. Therefore, the multi-label classification plays a leading role in the early optimization produce. Since WSDDN branch is directly supervised by image labels, the classification confidence produced by WSDDN will increase quickly. With the increment of classification confidence, the learning ability of the later branches will be also enhanced.

Optimization of Nesting Boxes Mining. Object detection is essentially learning a kind of ranking mechanisms, so that boxes with high classification probabilities will rank high. Therefore, the performance of a object detector is more relevant to the relative classification probabilities. This is the reason why NBM works well in boosting weakly supervised detection performance.

NBM can discover more positive samples, however it will also introduce false positive samples at the same time. Reviewing the mining strategy used in NBM, since NBM

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN [3]	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
ContextLocNet [13]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
OICR [2]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
PCL [23]	54.4	69.0	39.3	19.2	15.7	62.9	64.4	30.0	25.1	52.5	44.4	19.6	39.3	67.7	17.8	22.9	46.6	57.5	58.6	63.0	43.5
WRPN [24]	57.9	**70.5**	37.8	5.7	21.0	66.1	69.2	59.4	3.4	57.1	57.3	35.2	64.2	68.6	**32.8**	**28.6**	50.8	49.5	41.1	30.0	45.3
ONBR	**60.6**	65.9	**49.3**	**37.6**	**23.7**	**72.7**	67.8	50.0	21.1	**60.2**	43.5	**51.2**	61.5	67.3	13.7	24.3	48.9	**62.1**	**66.1**	**69.2**	**50.8**
OICR [2] + FRCNN	**65.5**	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	63.1	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
PCL [23] + FRCNN	63.2	69.9	47.9	22.6	27.3	71.0	69.1	49.6	12.0	60.1	51.5	37.3	63.3	63.9	15.8	23.6	48.8	55.3	61.2	62.1	48.8
WRPN [24] + FRCNN	63.0	69.7	40.8	11.6	**27.7**	70.5	**74.1**	58.5	10.0	**66.7**	**60.6**	34.7	**75.7**	**70.3**	25.7	26.5	**55.4**	56.4	55.5	54.9	50.4
W2F(FRCNN) [29]	64.0	67.4	49.9	32.8	15.0	71.8	69.2	70.6	24.2	55.2	49.2	64.9	54.3	65.3	24.3	23.0	49.6	60.1	60.0	62.8	51.7
W2F(FerRCNN) [29]	63.5	70.1	**50.5**	31.9	14.4	72.0	67.8	**73.7**	23.3	53.4	49.4	**65.9**	57.2	67.2	27.6	23.8	51.8	58.7	64.0	62.3	52.4
ONBR + FRCNN	60.7	60.8	43.6	32.7	24.2	63.5	68.2	70.1	25.7	58.7	47.5	60.2	66.1	66.0	21.5	23.0	50.0	52.4	65.8	68.6	51.5
ONBR + FerRCNN	65.1	66.4	50.4	**38.1**	24.3	66.9	70.9	61.4	**26.1**	66.3	47.5	61.9	67.4	68.4	20.5	23.3	52.8	52.1	64.9	68.4	**53.2**

Table 1. Mean Average Precition (in %) for different methods on VOC 2007 test set. The upper part shows the single model results, and the approaches shown in the lower part combine multiple models. "FRCNN" and "FerRCNN" indicates Fast R-CNN and Faster R-CNN respectively. Like-wise to later table.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	CorLoc
WSDDN [3]	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
ContextLocNet [13]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	47.1	59.5	67.0	83.5	35.3	39.9	67.0	49.7	63.5	65.2	55.1
OICR [2]	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	81.4	60.6
PCL [23]	79.6	**85.5**	62.2	47.9	37.0	83.8	83.4	43.0	38.3	80.1	50.6	30.9	57.8	90.8	27.0	58.2	75.3	**68.5**	75.7	78.9	62.7
WRPN [24]	77.5	81.2	55.3	19.7	44.3	80.2	86.6	69.5	10.1	**87.7**	68.4	52.1	84.4	91.6	**57.4**	**63.4**	77.3	58.1	57.0	53.8	63.8
ONBR	**86.7**	79.2	**70.6**	**60.6**	**46.9**	80.7	83.3	55.2	34.4	81.5	39.2	**66.7**	78.2	91.2	20.8	52.4	**78.4**	48.4	**83.3**	**82.1**	**66.0**
OICR [2] + FRCNN	85.8	82.7	62.8	45.2	43.5	**84.8**	**87.0**	46.8	15.7	82.2	51.0	45.6	83.7	91.2	22.2	59.7	75.3	65.1	76.8	78.1	64.3
PCL [23] + FRCNN	83.8	85.1	65.5	43.1	50.8	83.2	85.3	59.3	28.5	82.2	57.4	50.7	85.0	92.0	27.9	54.2	72.2	65.9	77.6	82.1	66.6
WRPN [24] + FRCNN	83.8	82.7	60.7	35.1	**53.8**	82.7	88.6	67.4	22.0	86.3	**68.8**	50.9	**90.8**	**93.6**	44.0	61.2	82.5	65.9	71.1	76.7	68.4
W2F(FRCNN) [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.4
W2F(FerRCNN) [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	70.3
ONBR + FRCNN	83.8	74.1	67.0	**61.2**	39.3	78.7	80.6	**75.9**	35.7	80.8	45.2	72.8	83.7	90.0	28.3	48.4	**84.5**	41.4	81.4	81.7	66.7
ONBR + FerRCNN	**89.6**	78.4	**72.4**	58.0	51.9	83.8	85.7	71.5	**45.6**	82.8	50.6	**74.0**	89.5	**93.6**	28.2	58.2	**84.5**	55.6	**84.8**	**82.4**	**71.1**

Table 2. Correct Localization (in %) for different methods on VOC 2007 trainval set.

considers the seed boxes and their inner and outer boxes as positive samples, the selected boxes most probably satisfy one of the following **three situations**:

1. if seed box is a tight box, then inner boxes are likely to be object parts, and outer boxes are likely to be oversized object.
2. if seed box is a partial box, then outer boxes are likely to be a tight object.
3. if seed box is a oversized box, then inner boxes are likely to be a tight object.

An additional analysis is illustrated in Figure. 3. We train a baseline multiple-instance detector on PASCAL VOC 2007 trainval dataset. We perform detection on the training set, and calculate the hitting rates of inner and outer boxes. From the above analysis we can find that, a tight box of an object have great change to be considered as positive sample and apply classification and regression. A object part will be considered as positive sample if seed box is a tight box or a partial box, however will be treat as background is seed box is a oversized box. Thus, the optimizing time of object parts will be lower than complete objects. Similar to oversized boxes. Therefore, after thousands of times of iteration, the classification confidence of tight boxes should higher than both partial and oversized boxes. So it is reasonable that NBM can boost the performance.

Optimization of Bounding Box Regressor. We adopt class agnostic strategy in bounding box regression, in which the regression parameters are shared over all categories. In the general scheme, simple samples are easier to classified correctly, thus will have greater confidence gain at the beginning of the training stage. The confidence also serves as the regression loss weight, so the regressor will first learn knowledge from simple samples. With the training ongoing, the knowledge learned from those simple classes will be progressively transferred to difficult classes.

4. Experiments

4.1. Experimental Setup

Datasets and evaluation metrics. We evaluate our approach on two widely used object detection benchmarks: PASCAL VOC 2007 and 2012, which have $9,962$ and $22,531$ images respectively. Both these two datasets contains 20 categories. The VOC 2007 involves $5,011$ images for training and the rest $4,952$ for test. The VOC 2012 contains $11,540$ images for training and the rest $10,991$ for test. We evade the boxes annotations in the dataset, and only use images and their label information for training. For evaluation metrics, we evaluate mean Average Precision (mAP) on test set, and we also adopt correct localization (CorLoc) on validation set to measure the localization accuracy [5]. Both the metrics are performed under the condition of $IoU > 0.5$

Implementation details. Implementation details. We adopt VGG16 as the base detection network [20]. We use

Method	mAP(%)	CorLoc(%)
OICR [2]	41.2	60.6
NBM	47.5	65.4
OICR + bbox [2]	46.3	64.2
ONBR (full model)	**50.8**	**66.7**

Table 3. Results of ablation study on VOC 2007 dataset.

selective search to generate about $2,000$ proposals for each image, which is same as [2]. We train the whole network end-to-end by using SGD with initial learning rate of 1×10^{-3}, weight decay of 0.0005 and momentum of 0.9. We perform $70,000$ iterations on PASCAL VOC dataset, and the learning rate will be divided by 10 at the $40,000$th iteration. We adopt the same multi scale settings in [2] to random select a scale in $\{480, 576, 688, 864, 1280\}$ at each iteration. Horizontal flip of all origin training images will be also added into training set. Besides, we also follow [2, 24] to use the detection results as pseudo ground truths to train another Fast R-CNN and Faster R-CNN. Such strategy always can further improve the performance. We perform NMS with threshold 0.3 on the detection results, and then select the boxes who have confidence score greater than 0.2. The setting of Fast R-CNN and Faster R-CNN are the same as in [8, 19].

4.2. Ablation Experiments

We conduct some ablation experiments to demonstrate the effectiveness of NBM and bounding box regressor. We adopt OICR as our base model, and perform experiments on VOC 2007 dataset as basline. We then plug our two propose modules in the base model under the same settings.

Influence of NBM. To validate the influence of NBM, we replace the positive samples selection strategy used in OICR as our NBM module. From Table. 3, we find that NBM boosts the performance from 41.2% to 47.5% on VOC 2007 test, and from 60.6% to 65.4% on VOC 2007 trainval. Because all the boxes come from selective search, thus the improvement indicates that NBM increase the classification confidence of correct instance boxes. This observation confirms the effectiveness of our NBM module.

Influence of Bounding Box Regression. We integrate a bounding box regressor into OICR, and find that this integration can also boost the performance significantly. We directly adopt the positive samples selected by the last refinement module as regression ground truths in the training stage. From Table. 3, bounding box regression can gain 5.1% for mAP (from 41.2% to 46.3%) and 3.6% for CorLoc (from 60.6% to 64.2%) on VOC 2007. Since bounding box regression does not change the classification score of each proposal, the performance gain indicates that such regression module do benefit to the detector, and makes localization more accurate.

Influence of ONBR. We further test the full setting of our ONBR. The regression targets are mined by the last NBM module. The mAP and CorLoc can increase to 50.8% and 66.7% respectively. The performance on each category could be found in Table. 1.

Influence of NBM. To validate the influence of NBM, we replace the positive samples selection strategy used in OICR as our NBM module. From Table. 3, we find that NBM boost the performance from 41.2% to 47.5% on VOC 2007 test, and from 60.6% to 65.4% on VOC 2007 trainval. Because all the boxes come from selective search, thus the improvement indicates that NBM increase the classification confidence of correct instance boxes. This phenomenon confirms our statement.

Influence of Bounding Box Regression. We integrate a bounding box regressor into OICR, and find that this integration can also boost the performance significantly. We directly adopt the positive samples selected by the last refinement module as regression ground truths in the training stage. From Table. 3, we can see that increase 5.1% mAP (from 41.2% to 46.3%) and 3.6% (from 60.6% to 64.2%) CorLoc on VOC 2007 dataset. Since bounding box regression does not change the classification score of each proposal, the performance gain indicates that such regression module do benefits to the detector, and makes the localization more accurate.

Influence of ONBR. We further test the full setting of our ONBR. The regression targets are mined by the last NBM module. The mAP and CorLoc can increase to 50.8% and 66.7% respectively. The detail performance on each category could be found in Table. 1.

4.3. Comparison with State-of-the-Art

We evaluate our approach on PASCAL VOC 2007 and 2012 dataset, report the performance and compare with state-of-the-arts [3, 13, 2]. The overall experiment results can be found in Table. 1, Table. 2, Table. 4, Table. 5.

The state-of-the-arts can be grouped into two kind of approach: Multi-Instance Learning(MIL) based approaches and pseudo-ground-truth based approaches. We first compare our proposed ONBR with MIL based approaches (i.e. the upper part of the tables). On VOC 2007 test dataset we achieve 0.8% mAP, which outperforms the state-of-the-arts[3, 13, 2] over 5.5% mAP absolutely. On VOC 2012 test dataset ONBR also achieve 46.1% mAP and outperforms others at least 5.3% mAP absolutely. ONBR is also robust to VOC 2007 and 2012 trainval datasets, which achieve 66.0% and 67.9% CorLoc on these two dataset respectively. From the four experiment tables we can find that ONBR beat all other single model.

Pseudo-ground-truth based approaches (i.e. the lower part of the tables) always first use MIL-based approaches to generate bounding boxes on training set, and then treat

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	CorLoc
OICR [2]	67.7	61.2	41.5	25.6	22.2	54.6	49.7	25.4	**19.9**	47.0	18.1	26.0	38.9	67.7	2.0	22.6	41.1	34.3	37.9	55.3	37.9
PCL [23]	58.2	66.0	41.8	24.8	27.2	55.7	55.2	28.5	16.6	51.0	17.5	28.6	49.7	70.5	7.1	25.7	47.5	36.6	44.1	59.2	40.6
WRPN [24]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.8
ONBR	**72.3**	**67.7**	**50.4**	**36.0**	**29.9**	**60.7**	**60.3**	**29.2**	18.0	**59.0**	**23.9**	**36.5**	**63.0**	**73.2**	5.4	**26.1**	**52.5**	**50.2**	**48.3**	**59.5**	**46.1**
OICR [2] + FRCNN	71.4	69.4	55.1	29.8	28.1	55.0	57.9	24.4	17.2	59.1	21.8	26.6	57.8	71.3	1.0	23.1	52.7	37.5	33.5	56.6	42.5
PCL [23] + FRCNN	69.0	**71.3**	**56.1**	30.3	27.3	55.2	57.6	30.1	8.6	56.6	18.4	43.9	64.6	71.8	7.5	23.0	46.0	44.1	42.6	58.8	44.2
WRPN [24] + FRCNN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.7
W2F(FRCNN) [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	47.3
W2F(FerRCNN) [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	47.8
ONBR + FRCNN	73.1	66.6	51.9	32.1	26.7	**61.8**	**60.7**	**56.1**	15.8	58.7	22.5	66.4	64.4	72.8	**9.9**	23.7	53.3	48.8	39.9	57.1	48.1
ONBR + FerRCNN	**74.1**	66.6	53.3	35.5	27.0	61.4	63.4	52.4	15.3	**62.3**	19.8	**66.9**	**67.0**	**73.2**	9.6	22.8	**54.9**	49.6	40.7	56.7	**48.6**

Table 4. Mean Average Precision (in %) for different methods on VOC 2012 test set.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	CorLoc
OICR [2]	86.2	84.2	68.7	55.4	46.5	82.8	74.9	32.2	46.7	82.8	42.9	41.0	68.1	89.6	9.2	53.9	81.0	52.9	59.5	83.2	62.1
PCL [23]	77.2	83.0	62.1	55.0	49.3	83.0	75.8	37.7	43.2	81.6	**46.8**	42.9	73.3	90.3	21.4	56.7	84.4	55.0	62.9	82.5	63.2
WRPN [24]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.9
ONBR	**89.9**	**86.3**	**72.3**	**63.8**	**55.9**	**88.1**	**79.9**	**38.2**	46.6	**86.7**	44.3	**53.3**	**88.2**	90.1	17.4	**61.8**	**87.1**	53.2	**70.0**	**84.2**	**67.9**
OICR [2] + FRCNN	89.3	86.3	**75.2**	57.9	53.5	84.0	79.5	35.2	**47.2**	87.4	43.4	43.8	77.0	91.0	10.4	60.7	86.8	55.7	62.0	**84.7**	65.6
PCL [23] + FRCNN	86.7	**86.7**	74.8	56.8	53.8	84.2	**80.1**	42.0	36.4	86.7	46.5	54.1	87.0	**92.7**	**24.6**	**62.0**	86.2	63.2	**70.9**	84.2	68.0
WRPN [24] + FRCNN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.3
W2F(FRCNN) [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.0
W2F(FerRCNN) [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.4
ONBR + FRCNN	**91.0**	82.7	70.4	**66.3**	48.6	88.1	79.3	58.9	41.5	**90.3**	45.1	73.9	**88.8**	91.0	22.1	58.3	88.7	67.5	66.2	79.8	69.9
ONBR + FerRCNN	90.0	83.2	70.5	**66.3**	48.0	**89.1**	79.9	**59.5**	41.9	89.6	46.3	**74.2**	88.0	91.2	22.5	57.5	**89.6**	**70.2**	64.9	81.2	**70.2**

Table 5. Correct localization (in %) for different methods on VOC 2012 trainval set.

the prediction as pseudo ground truth to train another fully-supervised detectors. Such operation always boosts the performance. ONBR is so powerful that its single model can even outperform most pseudo-ground-truth based approaches. From Table. 1 we can see that the result of ONBR single model is greater than the first three pseudo-ground-truth based approaches. For fair comparison, we also follow [] to use the output of ONBR to train a Fast R-CNN and a Faster R-CNN. As shown in the four tables, the model "ONBR + Faster R-CNN" outperforms all other approaches at least 1% mAP and 1% CorLoc absolutely, which proves the effectiveness of ONBR again.

By analysing the performance on each category, we find that ONBR is especially effective on animal-related categories (i.e. bird, cat, cow, dog, horse, sheep). The mAP on cat, dog and horse categories all increase more than 20% comparing with the OICR baseline model. The reason is that animals are not rigid objects, and may have more diverse poses. The experiment results confirm that our proposed nesting boxes selection can prevent the network from converging to a discriminate part.

4.4. Discussion

We visualize some results detected by ONBR intuitively in Figure. 4. The first two lines show the success cases, in which objects like dogs, sheep, etc could be correctly detected. From Table. 1 we observe that some categories, like person, does not perform well. We visualize some failure cases in the third row of Figure. 4 to analyze the failure reason. Our detector tends to find "head" instead of person. Because in some cases (e.g. 4^{th} image in 3^{rd} row), only a head appears in the image while the body is absent. Thus, the classifier in our network will learn the common pattern among images, this is why *head* are more likely to be discovered. We will study on this in the future work.

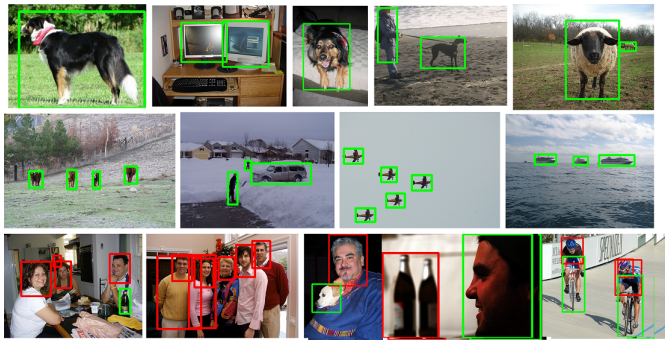

Figure 4. Some results produced by our approach (ONBR). Green boxes indicate the corrected prediction, while red ones indicate the failure cases.

5. Conclusion

In this paper, we propose a novel Online Nesting Boxes Regression (ONBR) network for weakly supervised object detection. In particular, we propose nesting boxes mining module to select positive samples online which are used to train instance classifiers. In addition, we integrate bounding box regressor into ONBR, which also adopt the pseudo ground truths selected by NBM as supervision. Because of the positive samples selection strategy, the network tends to converge to the complete objects. We conduct extensive experiments on PASCAL VOC 2007 and 2012 datasets to evaluate the performance of ONBR. The experiment results show our proposed ONBR achieves new state-of-the-arts.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, pages 577–584, 2003. 2

[2] P. T. X. W. X. Bai and W. Liu. Multiple instance detection network with online instance classifier refinement. 2017. 2, 3, 4, 6, 7, 8

[3] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 2, 3, 6, 7

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[5] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012. 6

[6] X. Dong, D. Meng, F. Ma, and Y. Yang. A dual-network progressive approach to weakly supervised object detection. In *ACMMM*, pages 279–287. ACM, 2017. 2

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 1

[8] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2, 5, 7

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 2

[10] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *PAMI*, 37(9):1904–1916, 2015. 2

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[12] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *CVPR*, pages 2883–2891, 2015. 2

[13] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, pages 350–365. Springer, 2016. 6, 7

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1

[15] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 1

[16] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, pages 3512–3520, 2016. 2

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1

[18] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *NIPS*, pages 570–576, 1998. 2

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 2, 7

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 6

[21] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *NIPS*, pages 1637–1645, 2014. 2

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. 1

[23] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. L. Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *PAMI*, 2018. 2, 6, 8

[24] P. Tang, X. Wang, A. Wang, Y. Yan, W. Liu, J. Huang, and A. Yuille. Weakly supervised region proposal network and object detection. In *ECCV*, pages 352–368, 2018. 6, 7, 8

[25] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *IJDWM*, 3(3):1–13, 2007. 2

[26] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 2, 3

[27] J. Wang, J. Yao, Y. Zhang, and R. Zhang. Collaborative learning for weakly supervised object detection. *arXiv:1802.03531*, 2018. 2

[28] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang. Ts2c: tight box mining with surrounding segmentation context for weakly supervised object detection. In *ECCV*, pages 454–470. Springer, Cham, 2018. 3

[29] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *CVPR*, pages 928–936, 2018. 2, 3, 6, 8

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2

[31] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Soft proposal networks for weakly supervised object localization. In *ICCV*, pages 1841–1850, 2017. 3

[32] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405. Springer, 2014. 2