

Cyclic-Bootstrap Labeling for Weakly Supervised Object Detection

Yufei Yin¹ Jiajun Deng² Wengang Zhou^{1,3,*} Li Li¹ Houqiang Li^{1,3,*}

¹ CAS Key Laboratory of Technology in GIPAS, EEIS Department, University of Science and Technology of China

² The University of Sydney

³ Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

yinyufei@mail.ustc.edu.cn, jiajun.deng@sydney.edu.au, {zhwg,lill,lihq}@ustc.edu.cn

Abstract

Recent progress in weakly supervised object detection is featured by a combination of multiple instance detection networks (MIDN) and ordinal online refinement. However, with only image-level annotation, MIDN inevitably assigns high scores to some unexpected region proposals when generating pseudo labels. These inaccurate high-scoring region proposals will mislead the training of subsequent refinement modules and thus hamper the detection performance. In this work, we explore how to ameliorate the quality of pseudo-labeling in MIDN. Formally, we devise Cyclic-Bootstrap Labeling (CBL), a novel weakly supervised object detection pipeline, which optimizes MIDN with rank information from a reliable teacher network. Specifically, we obtain this teacher network by introducing a weighted exponential moving average strategy to take advantage of various refinement modules. A novel class-specific ranking distillation algorithm is proposed to leverage the output of weighted ensembled teacher network for distilling MIDN with rank information. As a result, MIDN is guided to assign higher scores to accurate proposals among their neighboring ones, thus benefiting the subsequent pseudo labeling. Extensive experiments on the prevalent PASCAL VOC 2007 & 2012 and COCO datasets demonstrate the superior performance of our CBL framework. Code will be available at <https://github.com/Yinyf0804/WSOD-CBL/>.

1. Introduction

With the rapid advancements in deep neural networks, object detection has experienced significant progress. Nevertheless, state-of-the-art object detection methods are contingent upon accurate instance-level annotations obtained through fully-supervised learning settings. The process of collecting such annotations is both arduous and costly. Such

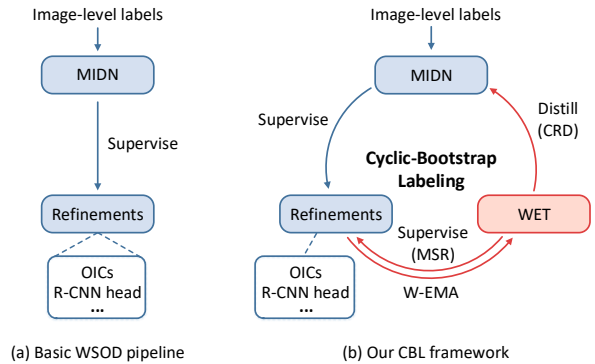


Figure 1. Comparison between the basic WSOD pipeline and our proposed CBL framework (feature extractor is omitted for simplicity). In the CBL framework, the subsequent refinement modules of MIDN are finally utilized to distill MIDN in turn, forming a cyclic-bootstrap procedure.

facts motivate the exploration of weakly supervised object detection (WSOD) [2, 32, 44, 43, 48, 25], which achieves the object detection task using only image-level labels.

Most existing WSOD approaches in the literature [32, 44] generally follow the training pipeline of Fig. 1(a). First, a multiple instance detection network (MIDN) [2] is obtained by leveraging multiple instance learning scheme for optimization, which converts WSOD into a multi-class classification problem over bottom-up generated region proposals [35]. The region proposals with high scores out of MIDN are then exploited to generate pseudo ground-truth boxes, which are used for the training of refinement modules, e.g., online instance classifiers (OICs), and regressors (R-CNN head). By introducing MIL to the training process, MIDN obtains the capability of estimating whether an object exists in the corresponding region. However, many high-scoring region proposals from MIDN only cover the discriminative part of an instance (e.g., the head of a bird), or contain some background regions [32, 48]. The improper scoring assignment of MIDN leads to the generation of inaccurate pseudo ground-truth boxes, which further hinder the training of subsequent refinement modules.

*Corresponding authors: Wengang Zhou and Houqiang Li

This problem has been recognized by the community, and several approaches have been proposed to address it. In the literature, C-MIDN [43] and P-MIDN [42] design complementary MIDN modules to find the remained discriminative parts other than the top-scoring ones. IM-CFB [46] develops a class feature bank to collect intra-class diversity information, and devises an FGIM algorithm to ameliorate the region proposal selection. These approaches primarily concentrate on the issue that the high-scoring proposals cover only the most discriminative parts, while struggling to address other inaccurate scoring-assignment issues. Besides, most of these attempts involve an auxiliary model, without offering assistance in the training of MIDN.

To this end, in this paper, we propose the Cyclic-Bootstrap Labeling (CBL) framework, which advances WSOD from the one-way pipeline (see Fig. 1 (a)) to a cyclic-bootstrap procedure (see Fig. 1 (b)). As shown in Figure, the subsequent modules of MIDN are eventually utilized to enhance itself. In comparison to the prior approaches, CBL exerts an additional rank-based supervision on MIDN, which is capable of handling a broader set of inaccurate scoring-assignment cases.

Specifically, following the common practice of WSOD, we first employ MIDN to generate pseudo labels, which serve as the initial supervision for subsequent refinement modules. To obtain more accurate classification results, we construct a weighted ensemble teacher (WET) model, inspired by the success of mean teacher methods [34, 23]. The WET model is updated through the weighted exponential moving average (W-EMA) strategy, which takes advantage of multiple student candidates in the refinement modules. Subsequently, leveraging the WET results, we propose a class-specific ranking distillation (CRD) algorithm to supervise MIDN with rank-based labels in a distillation manner. This additional rank-based supervision allows MIDN to achieve an improved scoring (ranking) assignment, where accurate proposals will be assigned higher scores among their neighboring ones. Moreover, we observe that the WET model can also act as a reliable teacher for the R-CNN head in the basic WSOD pipeline [44]. To this end, we propose a multi-seed R-CNN (MSR) algorithm to mine multiple positive seeds according to the WET results, calculate their confidence scores, and utilize them to generate pseudo labels for the supervision of the R-CNN head.

Our main contributions are summarized as follows:

- We propose a novel cyclic-bootstrap labeling (CBL) framework for weakly supervised object detection. The proposed CBL contains a weighted ensemble teacher model to generate reliable detection results, a class-specific ranking distillation algorithm to distill the MIDN module with rank information, and a multi-seed R-CNN algorithm to mine accurate positive seeds for the training of the R-CNN head.

- We provide a new perspective that the subsequent modules of MIDN are finally utilized to distill MIDN in turn, forming a cyclic-bootstrap procedure, which is rarely explored in previous WSOD works.
- Extensive experiments on the prevalent PASCAL VOC 2007 & 2012 and COCO datasets demonstrate the superior performance of our CBL framework.

2. Related Work

In this section, we briefly review the related methods including Weakly supervised object detection (WSOD) and Knowledge distillation.

2.1. Weakly Supervised Object Detection

Weakly supervised object detection (WSOD) [2, 17, 32, 31, 10, 28, 51, 40, 7, 19, 16, 33, 37, 1, 43, 36, 44, 48, 11, 21, 25, 4, 45, 46, 42, 49, 5, 50, 9, 27, 47, 15, 26, 20, 14, 30] has attracted much attention in recent years. Most recent works utilize Multiple Instance Learning (MIL) strategy to convert WSOD into a multi-class classification task, and adopt WSDDN [2] as the basic multiple instance detection network (MIDN) in their frameworks. WSDDN adopts MIL into a CNN network with a two-stream structure (*i.e.*, classification stream and detection stream), and combines the scores obtained from these two streams to generate instance-level scores. To improve the detection capability, on one hand, some works add several modules based on WSDDN for on-line refinement. OICR [32] first adds several cascaded on-line instance classifiers to refine the classification results, and adopts a top-scoring strategy to obtain pseudo seeds for training these classifiers. To obtain more accurate seeds, WSOD² [48] adopts bottom-up object evidence to update the original classification score during selection, [40] and [45] utilize results from the other tasks for assistance and MIST [25] proposes a multiple instance self-training algorithm. Furthermore, Yang [44] constructs a multi-task rnnn-head to adjust the positions and shapes of proposals.

On the other hand, some works propose to improve the basic MIDN. C-MIDN and P-MIDN [43, 42] design a (several) complementary MIDN module(s) to find the remained object parts other than the top-scoring one, WS-JDS [28] introduces the segmentation task for assistance, and IM-CFB [46] constructs a class feature bank to collect intra-class diversity information for amelioration. This work also aims to improve MIDN, but different from them, we propose to re-adjust the rank distribution of MIDN among neighboring positive instances, thus helping to generate more high-quality pseudo labels for subsequent refinement.

2.2. Knowledge Distillation

Knowledge distillation is firstly proposed in [13] to transfer knowledge from complicated teacher models to distilled student models, which makes the students achieve

similar performance to that of teachers. Knowledge distillation has been explored by a series of works in different tasks [3, 39, 12, 18, 41, 38]. [3] propose two new losses for better knowledge distillation on the classification and regression task, and combine hint learning to help the training process. [39] employ the inter-location discrepancy of teacher’s feature response on near object anchor locations for knowledge distillation. [12] use KL divergence loss to distill the classification head while processing proposals and negative proposals separately. [18] apply a similar rank distillation strategy with our CRD algorithm, while their main difference is that our work focuses on the WSOD task which does not have instance-level labels. To address this problem, we generate a reliable positive proposal set based on the more accurate WET results and propose a weighted KL divergence loss to alleviate the negative effects brought by noisy labels.

Previous WSOD work SoS [30] also adopts the knowledge distillation strategy. However, the usage in [30] simply follows the semi-supervised object detection paradigm and will not ameliorate the original WSOD network. In contrast, our method adopts data distillation in the WSOD training procedure to improve the rank distribution of MIDN, thus benefitting the whole WSOD network.

3. Our Method

The overall architecture of the proposed framework is shown in Fig. 2. An input image and a set of region proposals are first fed into the basic WSOD module. The proposal features are obtained through a CNN backbone and an RoI Pooling layer followed by two FC layers. Then, these features are fed into the MIDN module to produce instance-level scores. Meanwhile, the image and corresponding proposals are sent to the weighted ensemble teacher (WET) model, which is gradually updated by the basic WSOD module via a W-EMA strategy. After that, the WET results are utilized to distill the MIDN module with rank information through the class-specific ranking distillation (CRD) algorithm. Furthermore, WET acts a teacher to supervise the R-CNN head with the multi-seed R-CNN (MSR) algorithm.

3.1. Basic WSOD Module

Due to the lack of instance-level annotations in the WSOD settings, many existing works combine Multiple Instance Learning (MIL) with a CNN model, denoted as Multiple Instance Detection Network (MIDN), to accomplish the detection task. Following previous works [32, 31], we utilize a two-stream weakly supervised deep detection network (WSDDN) [2] as our MIDN module.

Given an image I , we denote its image-level label as $Y_{img} = [y_1, y_2, \dots, y_C] \in \mathbb{R}^{C \times 1}$, where $y_c = 1$ or 0 indicates the presence or absence of the class c . The generated region proposal set for image I is denoted as

$R = \{R_1, R_2, \dots, R_N\}$. We first extract proposal features through a CNN backbone and an RoI Pooling layer followed by two FC layers. Then, these proposal features are fed into two sub-branches in MIDN, *i.e.*, classification branch and detection branch. For classification branch, the score matrix $x^{cls} \in \mathbb{R}^{C \times |R|}$ is produced through a FC layer, where C and $|R|$ denote the number of categories and proposals, respectively. Then, a softmax operation is applied on x^{cls} along the categories to produce $\sigma_{cls}(x^{cls})$. Similarly, the score matrix $x^{det} \in \mathbb{R}^{C \times |R|}$ is produced in the detection branch by another FC layer. A softmax operation is then applied on x^{det} along proposals to produce $\sigma_{det}(x^{det})$. After that, the classification score for each proposal can be obtained by an element-wise product of these two scores: $x^{midn} = \sigma_{cls}(x^{cls}) \odot \sigma_{det}(x^{det})$. Finally, the image-level classification scores are generated through the summation over all proposals: $x_c^{img} = \sum_{i=1}^{|R|} x_{c,i}^{midn}$. In this way, we train the MIDN module with binary cross-entropy loss: $\mathcal{L}_{midn} = -\sum_{c=1}^C [y_c \log x_c^{img} + (1 - y_c) \log (1 - x_c^{img})]$

We further follow OICR [32] to add several cascaded on-line instance classifiers (OICs) to refine the classification results. Specifically, for each existing class, OICR selects the top-scoring proposal of the i -th classifier and its surrounding ones as positive samples to generate hard pseudo labels $Y_{ref_i} \in \mathbb{R}^{C+1 \times |R|}$. These labels are then used to train the subsequent $(i + 1)$ -th classifier using a weighted cross-entropy loss \mathcal{L}_{oic} . Particularly, pseudo labels of the first classifier OIC₁ are generated utilizing the MIDN scores.

Moreover, we add an R-CNN head following [44, 46], which consists of two parallel branches for the classification and regression task, respectively. The R-CNN head is supervised by the pseudo labels generated from the last on-line instance classifier. The weighted cross-entropy loss and smooth-L1 loss are applied for the two tasks, respectively.

3.2. Weighted Ensemble Teacher

In this section, we construct a sibling model of the basic WSOD module, Weighted Ensemble Teacher (WET), to produce more accurate detection predictions. Similar to the former, the WET model consists of a feature extractor and a classification head. An intuitive way to update the WET model is to apply Exponential Moving Average (EMA) following the traditional mean teacher methods [34, 23]:

$$\theta_t \leftarrow \alpha \theta_t + (1 - \alpha) \theta_s, \quad (1)$$

where θ_t and θ_s represent the parameters of the same network in the teacher model and the student model, respectively, and α is a smoothing coefficient. Through the EMA strategy, the slowly progressing teacher model can be considered as the ensemble of the student models in different training iterations [23]. We treat WET and the basic WSOD module as the teacher and student models, respectively.

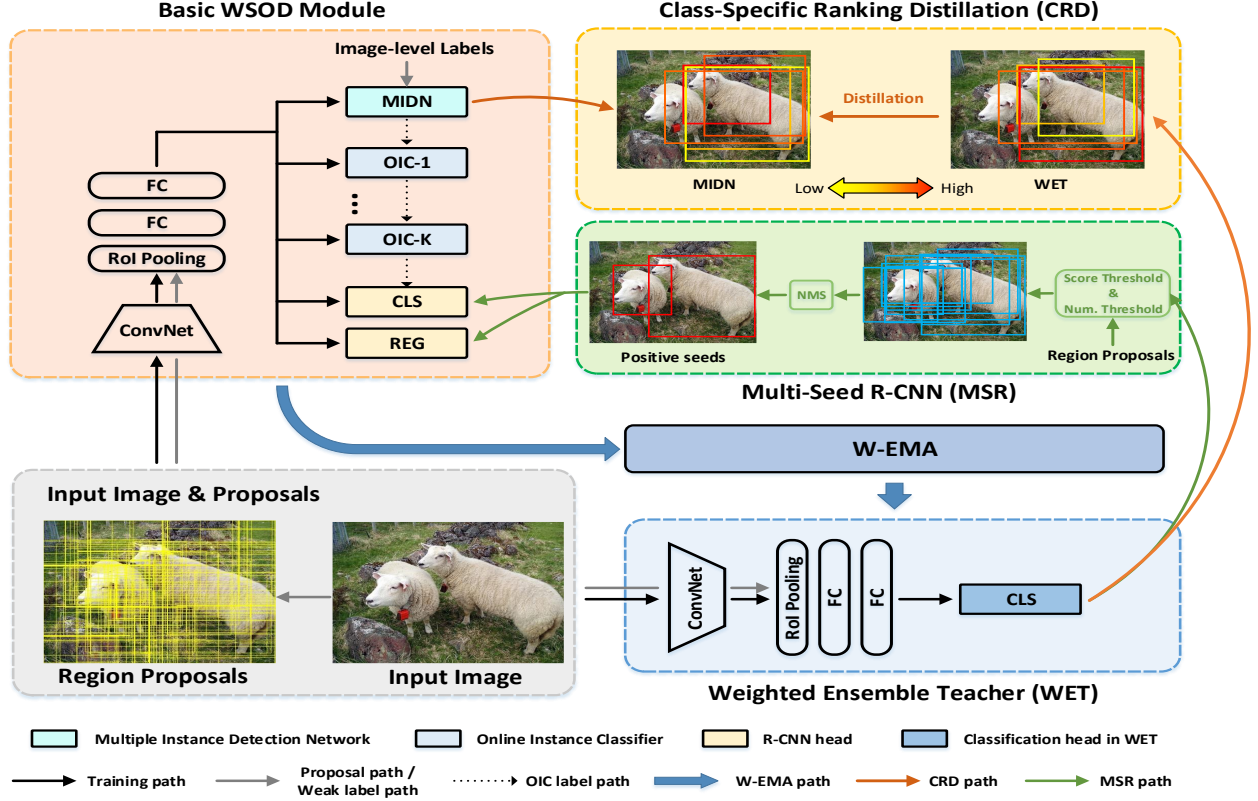


Figure 2. An overview of our CBL framework. Proposal features are first fed into the MIDN module to produce instance-level scores. Meanwhile, the image and corresponding proposals are sent to the weighted ensemble teacher (WET) model, which is gradually updated by the basic WSOD module via a W-EMA strategy. After that, the WET results are utilized to distill the MIDN module with rank information through the class-specific ranking distillation (CRD) algorithm. Furthermore, WET also acts as a teacher to supervise the R-CNN head with the multi-seed R-CNN (MSR) algorithm.

However, a problem arises that for the classification head in WET, many candidate networks can be treated as the **student** models, *i.e.*, K online instance classifiers (OICs) and classification branch (CLS branch) in R-CNN head. The most direct way is to choose CLS (or OIC_K) branch as the student network, considering the fact that they always have better performance among these candidates due to their relatively accurate pseudo labels after several refinements. However, this strategy overlooks the potential positive influence of the other candidate student networks.

To this end, we propose to modify EMA to accommodate multiple students. To be specific, we can use the average parameters of all these candidate student networks during EMA (A-EMA) instead of selecting a single candidate:

$$\theta_t \leftarrow \alpha \theta_t + \frac{(1-\alpha)}{S} \sum_{s=1}^S \theta_s, \quad (2)$$

where $S = K+1$ represents the number of candidates. Nevertheless, assigning the same weight to all the candidates is not the most efficient strategy due to the discrepancies in their performance. To enable the classification head to benefit more from the better candidate, we devise a weighted

EMA (W-EMA) to adjust the weight of these candidates accordingly:

$$\theta_t \leftarrow \alpha \theta_t + \frac{(1-\alpha)}{2} \left(\frac{1}{K} \sum_{k=1}^K \theta_k + \theta_{cls} \right), \quad (3)$$

where θ_k and θ_{cls} represent the parameters of k -th OIC and CLS branch, respectively. As a result, the weight of CLS branch is amplified to $\frac{K+1}{2}$ of its original value, while the weights of other candidates are decreased. It is noteworthy that W-EMA does not add extra hyperparameters, since it can be viewed as a two-step average of different student parameters (1st for OICs, 2nd for OIC-avg & CLS).

Overall, we employ EMA for updating feature extractor and W-EMA for updating classification head. In this paper, we refer to them collectively as W-EMA strategy. The WET model with W-EMA strategy has two main advantages: First, it can reduce the adverse effects of noisy pseudo labels. Second, the WET model can be regarded as an ensemble model of different student models at different time steps. These advantages enable the WET model to generate more reliable classification results $x^{wet} \in \mathbb{R}^{(C+1) \times |R|}$.

3.3. Class-Specific Ranking Distillation

Given the image-level supervision, MIDN is likely to assign high scores to some erroneous region proposals, such as detecting only the most discriminative parts or containing background noises. To alleviate this problem, we propose to employ additional supervision on the MIDN module. An intuitive way is to generate **classification-based** supervision with hard pseudo labels, similar to those used for refinement modules [32, 31]. However, this will exceed the limits of the original MIL constraint and, more importantly, contradict our original goal of solving the inaccurate scoring assignment problem of MIDN. For further discussions please refer to Supplementary Material.

To this end, instead of applying supervision for each individual proposal, we turn to distill MIDN with the **rank-based** information among associated proposals. The most straightforward approach for rank distillation is to drive MIDN to generate a rank distribution similar to that of WET for *all proposals*. However, this approach has two main drawbacks: On one hand, learning the rank distribution among inaccurate samples or irrelevant samples is futile. On the other hand, without instance-level annotations, some positive samples will inevitably be assigned lower scores than some negative ones. To ameliorate these issues, we design a Class-Specific Ranking Distillation (CRD) algorithm to guide MIDN to adjust to a more appropriate rank distribution among *confident associated proposals*.

Specifically, for an existing object class c (i.e., $y_c = 1$), we first select the proposal with the highest WET score on this class R_{i_c} , which is the most confident positive sample. Then, we calculate the overlaps between all proposals with the top-scoring one R_{i_c} , and set an overlap threshold τ to construct a neighboring positive proposal set P_c :

$$P_c = \{R_i | IoU(R_i, R_{i_c}) > \tau, R_i \in R\}. \quad (4)$$

Furthermore, to encourage MIDN to focus on the rank distribution under different views, we continuously increase the overlap threshold τ with a linear growth strategy:

$$\tau = \tau_0 + (\tau_1 - \tau_0) \frac{iter_{cur}}{iter_{max}}, \quad \tau \in [\tau_0, \tau_1], \quad (5)$$

where $iter_{cur}$ represents the current iteration, and $iter_{max}$ represents the total training iteration using CRD algorithm. τ_0 and τ_1 are set to 0.5 and 1.0 naturally following the common evaluation metrics for selecting positive proposals.

After obtaining the positive proposal set P_c for class c , we use their predicted scores to represent the rank distribution, since a higher score implies that the corresponding proposal will receive a higher rank in a particular class. We opt for the soft score instead of a hard ranking number since the soft supervision target in distillation is more effective in preserving detailed rank information. Then, a softmax operation is applied on their c -th scores for normalization to

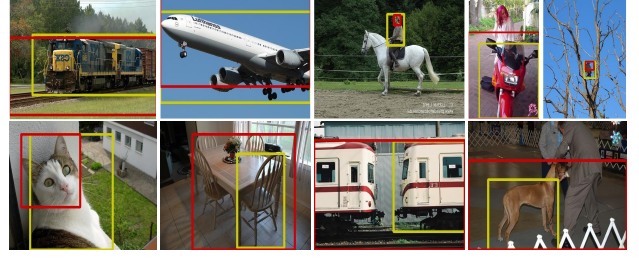


Figure 3. Comparison of the top-scoring proposals of the MIDN module in different frameworks at 20k iteration. The proposals from basic WSOD framework are in red and proposals from our CBL framework are in yellow.

represent the rank distribution for class c :

$$s'_{c,j} = \frac{e^{x_{cj}^{midn}}}{\sum_{k=1}^{|P_c|} e^{x_{ck}^{midn}}}, \quad t'_{c,j} = \frac{e^{x_{cj}^{wet}}}{\sum_{k=1}^{|P_c|} e^{x_{ck}^{wet}}}, \quad R_j, R_k \in P_c, \quad (6)$$

where s'_c and t'_c represent the rank distribution of MIDN (student) and WET (teacher) for class c .

Finally, we utilize a weighted KL divergence loss to distill the MIDN with rank distributions from WET:

$$\mathcal{L}_{crd} = - \sum_c \mathbb{I}(y_c = 1) \frac{w_c}{|P_c|} \sum_{j=1}^{|P_c|} t'_{c,j} \log\left(\frac{s'_{c,j}}{t'_{c,j}}\right), \quad (7)$$

where w_c represents the loss weight of class c . We apply the highest WET score on this class ($x_{ci_c}^{wet}$) as w_c to represent the confidence of the selected proposal set of this class (P_c).

By utilizing the CRD algorithm, MIDN is encouraged to adjust higher ranks to more accurate proposals compared with that in the original framework, as shown in Fig. 3.

3.4. Multi-Seed R-CNN

Due to its good performance, WET model can serve as a reliable teacher to other networks in the basic WSOD module apart from MIDN. In this section, we propose to employ WET for the supervision of the R-CNN head.

The pseudo label generation for R-CNN head in the original WSOD module can be divided into two steps: First, the top-scoring proposal (scores are from the last online instance classifier OIC_K) for an existing class is selected as the positive seed for this class. Then, pseudo labels of all proposals are generated according to the overlaps with the positive seed. This procedure guarantees the quality of the selected seeds, yet overlooks the potential advantages from other possible seeds in the same image.

To this end, we propose a simple Multi-Seed R-CNN (MSR) algorithm to generate more credible seeds by leveraging the reliable WET results. First, we use the ensemble of the WET results with the results from the original teacher OIC_K : $x^{msr} = (x^{wet} + x^{OIC_K})/2$. Next, we propose to narrow the search range of positive seeds. Taking into account the fluctuating distribution of scores during the train-

ing stage, we set a soft threshold to accomplish this. To be specific, for each existing class c , the threshold σ_c^s is denoted as the highest x_c^{msr} of this class multiplied with a factor μ^s . To avoid leaving too many proposals, we further set a threshold σ^n to limit the number of remaining proposals. σ^n is denoted as the number of whole proposals multiplied with a factor μ^n :

$$\sigma_c^s = \mu^s \max x_c^{msr}, \quad \sigma^n = \mu^n |R|. \quad (8)$$

We first select the top- σ^n proposals and then filter out the proposals with scores lower than σ_c^s . Then, we apply the Non-Maximum Suppression (NMS) algorithm to the remained proposals and regard the kept ones as positive seeds.

To further reduce the impact of noisy seeds, we apply the original results (*i.e.*, OIC_K and WET scores) as references to assess their confidence. Specifically, according to each result, we first remove negative proposals using Eq. 8. Then, for each seed i , we identify if there is one that is very close to it in the remaining proposals. After that, we calculate the proportion $p_{i,c}$ of such case for seed i according to all the results. A high proportion indicates that the seed is recognized as “positive” by multiple classifiers, thus making it more confident. Finally, the confidence of a seed is obtained as follows:

$$w_i = x_{i,c}^{msr} \cdot (1 + p_{i,c}^\gamma). \quad (9)$$

We generate corresponding pseudo labels according to these positive seeds for the classification branch and regression branch in R-CNN head. We apply weighted cross-entropy loss and weighted smooth-L1 loss to train these two branches, respectively. The confidence w_i is used as the weight and the loss for the R-CNN head \mathcal{L}_{rcnn} is obtained by combining these two losses. For more details, please refer to Supplementary Material.

3.5. Training Objectives

The overall training objective is the combination of MIDN, online instance classifiers (OICs), and R-CNN head, which is elaborated as follows:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{midn} + (1 - \lambda) \mathcal{L}_{crd} + \sum_{k=1}^K \mathcal{L}_{oic} + \mathcal{L}_{rcnn}, \quad (10)$$

where λ controls the weight of image-level supervision \mathcal{L}_{midn} and ranking distillation \mathcal{L}_{crd} . At the start of training process, MIDN needs to focus on the basic MIL learning for better multi-class classification, while gradually transitioning its focus to adjusting the rank distribution as the training progresses. To this end, we apply a linear decay strategy to adjust λ from 1 to 0 with the increment of training iteration. Additionally, we start using MSR algorithm at the $0.4 \cdot \text{maxiter}$ iteration, considering WET is still in the initial update stage at the beginning of the training procedure.

Methods	mAP@0.5	mAP@[.5, .95]
PCL [31]	19.4	8.5
MIST [25]	24.3	11.4
CASD [15]	26.4	12.8
Ours	27.6	13.6

Table 1. Performance comparison among the state-of-the-art methods with single model on MSCOCO dataset.

4. Experiments and Analysis

4.1. Datasets

We evaluate our method on the prevalent Pascal VOC 2007, Pascal VOC 2012 [8] and MSCOCO [22] datasets. For VOC 2007 & 2012 datasets, we train on *trainval* split (5,011 and 11,540 images for VOC 2007 & 2012), and report the average precision (AP) on *test* set, together with the correct localization rate (CorLoc) on *trainval* set. Only when the Jaccard overlap between the predicted bounding box and the corresponding ground-truth box is above 0.5, the prediction is regarded as a true positive one. For MSCOCO dataset, we train on the *train* split (82,738 images), and test on its *val* split (4,000 images). During evaluation, we apply two metrics mAP@0.5 and mAP@[.5, .95] following the standard MSCOCO criteria, respectively.

4.2. Implementation Details

We follow the common practice [32, 31] to exploit VGG16 [29] pre-trained on Imagenet [6] as the backbone network, and to apply Selective Search [35] for region proposal generation. We use SGD for optimization, and momentum and weight decay are set to 0.9 and 5×10^{-4} respectively. The learning rate is set to 1×10^{-3} for the first 50K iterations and 1×10^{-4} for the following 20K iterations. We set $\alpha = 0.999$, $\gamma = 0.4$, $\mu^s = 0.7$ and $\mu^n = 0.05$, and K is set to 3 following the common practice. iter_{max} is set to 80k for VOC 2007 & 2012 datasets. Following previous works [32, 25], multi-level scaling and horizontal flipping data augmentation are conducted in both training and testing. Our method is implemented on PyTorch [24], and we run all the experiments on an NVIDIA GTX 1080Ti GPU with a batch size of 4.

4.3. Comparison with State-of-the-arts

In Tab. 2, we present a comprehensive comparison of our proposed method with existing arts with single model on the VOC 2007 dataset. Our method achieves state-of-art performances of 57.4% mAP and 71.8% CorLoc, surpassing previous methods by at least 0.6% and 0.8%. Our method outperforms recent works [44, 25] that directly use original MIDN module to train cascaded refinement modules, since the proposed CRD algorithm improves the scoring assignment on the valuable neighboring positive proposals in MIDN, which benefits the subsequent pseudo la-

Methods	mAP (%)	CorLoc (%)
OICR [32]	41.2	60.6
PCL [31]	43.5	62.7
C-MIL [36]	50.5	65.0
Yang <i>et al.</i> [44]	51.5	68.0
C-MIDN [43]	52.6	68.7
SLV [4]	53.5	71.0
WSOD ² [48]	53.6	69.5
IM-CFB [46]	54.3	70.7
MIST [25]	54.9	68.8
CASD [15]	56.8	70.4
Ours	57.4	71.8

Table 2. Performance comparison among the state-of-the-art methods with single model on PASCAL VOC 2007.

beling, thus raising the upper limit of the whole WSOD performance. Some methods [43, 46] also aim to deal with the inaccurate scoring assignment issues of MIDN, but their focus is primarily on the part domination problem that high-scoring proposals surround only the discriminative parts. Different from them, our method distills MIDN with rank information from a reliable WET model, which guides MIDN to assign higher scores to accurate proposals among their neighboring ones. Hence, our work can also handle other cases with inaccurate high-scoring proposals (*e.g.*, containing background noises). Furthermore, our method makes the whole framework as a cyclic-bootstrap procedure through the model ensemble (W-EMA strategy) and the rank-information distillation (CRD algorithm). Therefore, our method also performs better than them.

For the MSCOCO dataset, as shown in Tab. 1, our method produces the state-of-art performances of 27.6% mAP@0.5 and 13.6% mAP@[.5, .95], outperforming the best competitor CASD [15] by clear margins of 1.2% and 0.8%, which also validates the effectiveness of our work.

Fig. 4 shows the detection results on VOC 2007. The first two rows indicate that our method can detect multiple instances accurately (*e.g.*, “dog”, “car”), even if they are in some complex scenes. Some failure cases are shown in the last row, which contains localizing only the discriminative parts (*e.g.*, human faces), grouping several objects (especially for “bottle” class), and containing background parts.

4.4. Ablation Study

4.4.1 Effect of Each Component

We conduct ablation studies on the main components of CBL framework in Tab. 3 under the mAP metric, where “inf.” represents using WET model during inference. We start from the basic WSOD module (Line 1), with an mAP of 53.3%. Next, we extend the basic model by adding WET model and use it to distill the MIDN module with CRD algorithm (Line 3), which improves the basic model to 55.8% mAP, bringing a clear 2.5% gain. This outcome highlights the effectiveness of the CRD algorithm in enhancing the

Basic	WET	CRD	MSR	Inf.	mAP (%)
✓					53.3
✓		✓			54.2
✓	✓	✓			55.8
✓	✓	✓		✓	56.0
✓	✓	✓	✓	✓	57.4

Table 3. Ablative experiments on the effects of different components in our CBL. The models are evaluated on PASCAL VOC 2007 in terms of mAP (%).

Updating Strategy	mAP (%)
EMA with the last OIC branch	55.3
EMA with the CLS branch	54.5
A-EMA with all OICs and CLS branch	56.1
W-EMA with all OICs and CLS branch	57.4

Table 4. Ablative experiments on the effect of the WET updating strategy. The models are evaluated on PASCAL VOC 2007.

Overlap Threshold	mAP (%)
Static Value ($\tau = 0.75$)	55.8
Linear Growth ($\tau \in [0.5, 1.0]$)	57.4
Linear Decline ($\tau \in [0.5, 1.0]$)	56.2

Table 5. Ablative experiments on the effect of overlap threshold in CRD. The models are evaluated on PASCAL VOC 2007.

overall WSOD performance by distilling rank information on MIDN. After that, we utilize the WET model during inference with the proposed weighted ensemble strategy (Line 4), boosting the performance to 56.0 % mAP.

To validate the importance of the proposed WET model, we conduct an additional experiment where we replace the WET model with the branch in the basic model (*i.e.* the last OIC branch OIC_K or the classification branch in R-CNN head) for distillation (Line 2), and we find OIC_K performs better with an mAP of 54.2%. However, this operation resulted in a 1.8% mAP drop, indicating that the WET model is a more dependable teacher in the distillation process. Nonetheless, the performance still remains 0.9% superior to the basic model, thus further validating the efficacy of the CRD algorithm. Moreover, when applying the MSR algorithm (Line 5), we can achieve the best performance 57.4%, which shows that WET can also function as a proficient teacher during the training of the R-CNN head.

4.4.2 Effect of WET Updating Strategy

We conduct experiments to analyze the influence of different updating strategies on WET. The results are shown in Tab. 4, where OIC and CLS represent the online instance classification branch and the classification branch in R-CNN head, respectively. When directly using the single classification branch to update the WET model via EMA (Line 1-2), the proposed WET model can achieve at most 55.3% mAP, which demonstrates the effectiveness of the whole CBL framework. In addition, utilizing CLS branch

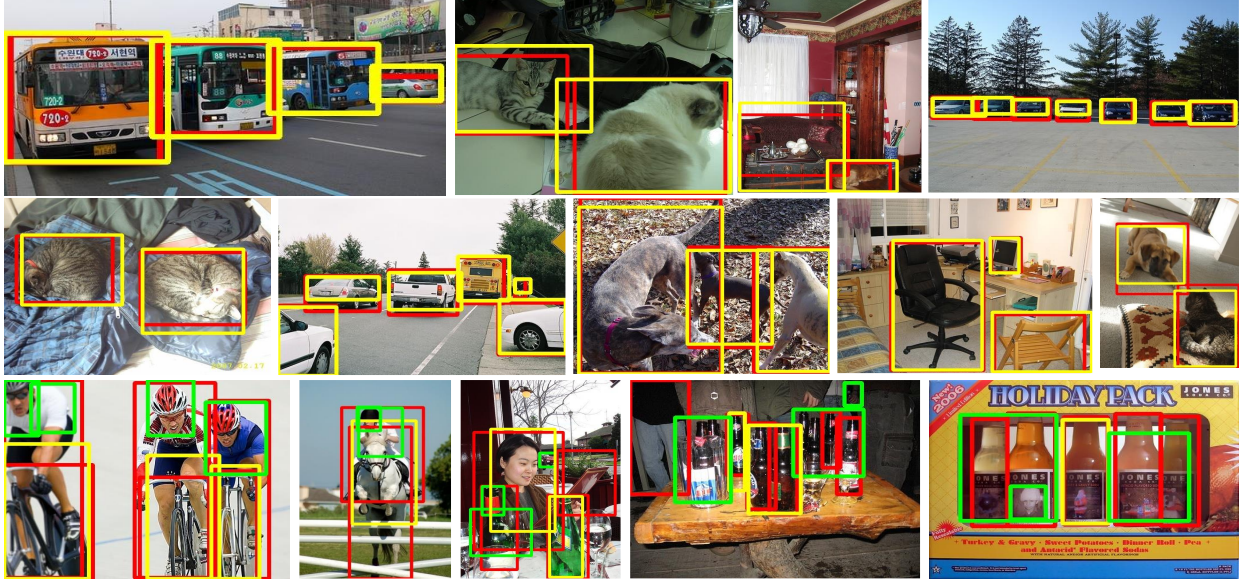


Figure 4. Visualization results on VOC 2007 *test* set. Boxes in red, yellow, and green represent ground-truth boxes, successful predictions, and failure cases, respectively.

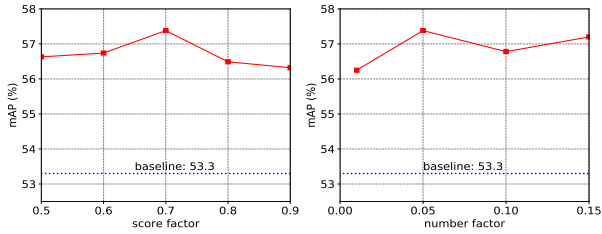


Figure 5. Influences of score factor μ^s and number factor μ^n in MSR. The models are evaluated on PASCAL VOC 2007.

does not perform well since it is cascaded after too many refinement modules, hence converging slowly at the beginning of the training procedure. Directly employing it to update WET will influence the critical initial update phase of WET. When A-EMA is applied (Line 3), the performance shows an improvement of 0.8% mAP, indicating that other candidates, apart from the best ones, can also have a positive impact on updating the WET model. Additionally, when using the W-EMA strategy (Line 4), the whole framework achieve the best performance 57.4% mAP. This result shows that assigning a higher weight to the superior candidate during the update process is a more effective strategy.

4.4.3 Overlap Threshold in CRD

We compare the different settings on the overlap threshold τ in CRD, and the results are presented in Tab. 5. We find that changing τ during training brings more benefits than directly setting a static value, since the former setting will help MIDN to pay attention to the rank distribution under different views. Moreover, a linear growth strategy performs best, because the pseudo labels for the subsequent refinement module need to be more precise with its increasing detection capability. Therefore, CRD algorithm needs to

gradually narrow the view to focus on fine-tuning the rank distribution of more accurate proposals.

4.4.4 Effect of MSR Selection Range

Fig. 5 shows the influences of score factor μ^s and number factor μ^n used to narrow the selection range of positive seeds. Among all the settings, $\mu^s = 0.7, \mu^n = 0.05$ performs best. If the range is too small (large μ^s or small μ^n), few seeds will be found, which limits the benefits from MSR algorithm. Conversely, if the range is too large, some noisy samples will be selected incorrectly, thus degrading the MSR performance. Our MSR algorithm is insensitive to both μ^s and μ^n and all the settings outperform the baseline by at least 3.0% mAP.

5. Conclusion

In this paper, we propose an effective cyclic-bootstrap labeling (CBL) framework for WSOD. We first construct a reliable WET model and update it via W-EMA strategy. After that, the WET results are utilized to distill the MIDN module with rank distribution with the proposed CRD algorithm. Additionally, we propose an MSR algorithm to mine accurate positive seeds to train the R-CNN head better. The whole framework acts as a cyclic-bootstrap procedure where the subsequent modules of MIDN are finally utilized to supervise itself. Extensive experiments on the PASCAL VOC 2007 & 2012, and MSCOCO datasets demonstrate the superior performance of our CBL framework.

Acknowledgement: This work was supported by NSFC under Contract U20A20183 and 62021001. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and the Supercomputing Center of the USTC.

References

- [1] Aditya Arun, CV Jawahar, and M Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9432–9441, 2019.
- [2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016.
- [3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.
- [4] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Slv: Spatial likelihood voting for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12995–13004, 2020.
- [5] Gong Cheng, Junyu Yang, Decheng Gao, Lei Guo, and Junwei Han. High-quality proposals for weakly supervised object detection. *IEEE Transactions on Image Processing*, 29:5794–5804, 2020.
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
- [7] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 914–922, 2017.
- [8] 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 (IJCV)*, 88(2):303–338, 2010.
- [9] Xiaoxu Feng, Junwei Han, Xiwen Yao, and Gong Cheng. Tcanet: Triple context-aware network for weakly supervised object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6946–6955, 2020.
- [10] Mingfei Gao, Ang Li, Ruichi Yu, Vlad I Morariu, and Larry S Davis. C-wsl: Count-guided weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 152–168, 2018.
- [11] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. Utilizing the instability in weakly supervised object detection. *arXiv preprint arXiv:1906.06023*, 2019.
- [12] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021.
- [13] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.
- [14] Zitong Huang, Yiping Bao, Bowen Dong, Erjin Zhou, and Wangmeng Zuo. W2n: Switching from weak supervision to noisy supervision for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 708–724. Springer, 2022.
- [15] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly supervised object detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:16797–16807, 2020.
- [16] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1377–1385, 2017.
- [17] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 350–365, 2016.
- [18] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 2022.
- [19] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9735–9744, 2019.
- [20] Mingxiang Liao, Fang Wan, Yuan Yao, Zhenjun Han, Jialing Zou, Yuze Wang, Bailan Feng, Peng Yuan, and Qixiang Ye. End-to-end weakly supervised object detection with sparse proposal evolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 210–226. Springer, 2022.
- [21] Chenhao Lin, Siwen Wang, Dongqi Xu, Yu Lu, and Wayne Zhang. Object instance mining for weakly supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11482–11489, 2020.
- [22] 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, 2014.
- [23] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021.
- [24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [25] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10598–10607, 2020.

- [26] Jinhwan Seo, Wonho Bae, Danica J Sutherland, Junhyug Noh, and Daijin Kim. Object discovery via contrastive learning for weakly supervised object detection. In *European conference on computer vision*, pages 312–329. Springer, 2022.
- [27] Yunhang Shen, Rongrong Ji, Yan Wang, Zhiwei Chen, Feng Zheng, Feiyue Huang, and Yunsheng Wu. Enabling deep residual networks for weakly supervised object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 118–136. Springer, 2020.
- [28] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Lijuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 697–707, 2019.
- [29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [30] Lin Sui, Chen-Lin Zhang, and Jianxin Wu. Salvage of supervision in weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14227–14236, 2022.
- [31] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Loddon Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(1):176–191, 2018.
- [32] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2843–2851, 2017.
- [33] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–368, 2018.
- [34] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.
- [35] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013.
- [36] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2199–2208, 2019.
- [37] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1306, 2018.
- [38] Kuo Wang, Jingyu Zhuang, Guanbin Li, Chaowei Fang, Lechao Cheng, Liang Lin, and Fan Zhou. De-biased teacher: Rethinking iou matching for semi-supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2573–2580, 2023.
- [39] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019.
- [40] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 434–450, 2018.
- [41] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.
- [42] Yunqiu Xu, Chunlun Zhou, Xin Yu, Bin Xiao, and Yi Yang. Pyramidal multiple instance detection network with mask guided self-correction for weakly supervised object detection. *IEEE Transactions on Image Processing (TIP)*, 30:3029–3040, 2021.
- [43] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9833–9842, 2019.
- [44] 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 (ICCV)*, pages 8372–8381, 2019.
- [45] Ke Yang, Peng Zhang, Peng Qiao, Zhiyuan Wang, Dongsheng Li, and Yong Dou. Objectness consistent representation for weakly supervised object detection. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pages 1688–1696, 2020.
- [46] Yufei Yin, Jiajun Deng, Wengang Zhou, and Houqiang Li. Instance mining with class feature banks for weakly supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 3190–3198, 2021.
- [47] Yufei Yin, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. Fi-wsod: Foreground information guided weakly supervised object detection. *IEEE Transactions on Multimedia*, 2022.
- [48] 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 International Conference on Computer Vision (ICCV)*, pages 8292–8300, 2019.
- [49] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.
- [50] Dingwen Zhang, Wenyuan Zeng, Jieru Yao, and Junwei Han. Weakly supervised object detection using proposal-and semantic-level relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

- [51] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4262–4270, 2018.