W2N: Switching From Weak Supervision to Noisy Supervision for Object Detection

Zitong Huang[1], Yiping Bao[2], Bowen Dong[1], Erjin Zhou[2], and Wangmeng Zuo[1,3]✉

[1]Harbin Institute of Technology [2]MEGVII Technology
[3]Peng Cheng Laboratory
{zitonghuang99,cndongsky}@gmail.com,{baoyiping,zej}@megvii.com,
wmzuo@hit.edu.cn

Abstract. Weakly-supervised object detection (WSOD) aims to train an object detector only requiring the image-level annotations. Recently, some works have managed to select the accurate boxes generated from a well-trained WSOD network to supervise a semi-supervised detection framework for better performance. However, these approaches simply divide the training set into labeled and unlabeled sets according to the image-level criteria, such that sufficient mislabeled or wrongly localized box predictions are chosen as pseudo ground-truths, resulting in a sub-optimal solution of detection performance. To overcome this issue, we propose a novel WSOD framework with a new paradigm that switches from weak supervision to noisy supervision (W2N). Generally, with given pseudo ground-truths generated from the well-trained WSOD network, we propose a two-module iterative training algorithm to refine pseudo labels and supervise better object detector progressively. In the localization adaptation module, we propose a regularization loss to reduce the proportion of discriminative parts in original pseudo ground-truths, obtaining better pseudo ground-truths for further training. In the semi-supervised module, we propose a two tasks instance-level split method to select high-quality labels for training a semi-supervised detector. Experimental results on different benchmarks verify the effectiveness of W2N, and our W2N outperforms all existing pure WSOD methods and transfer learning methods. Our code is publicly available at https://github.com/1170300714/w2n_wsod.

Keywords: weakly supervised learning, object detection

1 Introduction

Different from fully supervised object detection (FSOD) [10,24] which heavily relys on instance-level bounding box annotations, weakly supervised object detection (WSOD) aims to use only image-level labels as supervision to train an object detector. Compared to the time-consuming instance-level ground-truth annotating process, image-level category labels are easy to obtain relatively,

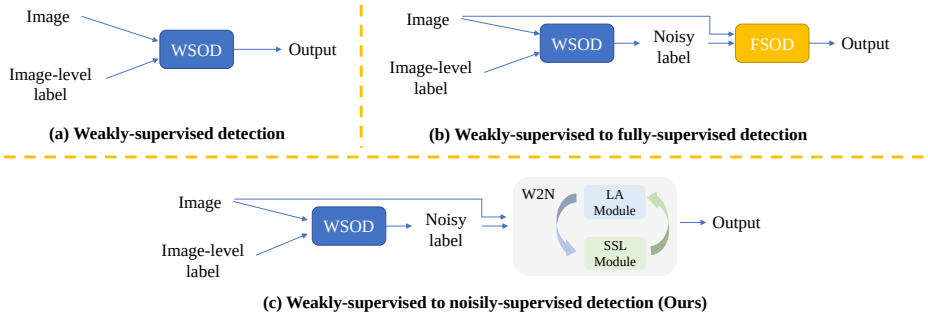

Fig. 1: Training paradigms with three different weakly supervised object detection frameworks: (a) Basic weakly-supervised detection. (b) Weakly-supervised to fully-supervised detection framework. (c) Our W2N framework.

which is more time-saving and economy. Therefore, WSOD has become a hot and meaningful research topic. Existing WSOD methods [2,31,30,25,5] usually follow the multiple instance learning (MIL) framework, which is based on pre-computed region proposals [33] and is formulated as a proposals classification task, as shown in Fig. 1 (a). However, without accurate bounding box ground-truths, the localization ability of model is severely limited by inaccurate region proposals. Specifically, the WSOD network tends to focus on the discriminative part instead of the whole object for some typical categories (person, cat, dog, etc.). As shown in Fig. 1 (b), some works [41,15,31,30,35,38] proposed pseudo ground-truth (PGT) excavation algorithm to generate pseudo ground-truths from prediction by a MIL-based weakly-supervised object detector and use it to deploy a supervised detector, trying to apply the FSOD training paradigm to WSOD task. However, the improvement of detection precision is still limited because some low-quality boxes in the pseudo ground-truths make the WSOD network converge to the sub-optimal solution.

To reduce the negative effect from low-quality pseudo ground-truths, some semi-supervised learning [22,28] approaches have been proposed and applied into weakly supervised object detection tasks. *e.g.*, the recently proposed SoS [29] combines a novel labeled-unlabeled dataset split method as well as the state-of-the-art semi-supervised detection method [22] into the WSOD training to improve the detection performance. The main idea of this method is paying more attention to relatively high-quality pseudo labels and carry out a dynamic label updating for noisy labels to improve the performance of detector progressively.

Inspired by this semi-supervised learning formula, we argue that the pseudo ground-truths can been seen as an inaccurate instance-level bounding box annotation, so it's significant to formulate the multi-phase WSOD problem as a noisy-label object detection task. To this end, we propose our novel weakly supervised object detection framework namely Weakly-supervision to Noisy-supervision (W2N). The noisy labels of the training image set are generated by

any well-trained WSOD and then fed into W2N framework for further training procedure. An overview of the contrast between the existing WSOD framework and our framework is presented in Fig. 1 (c).

We formulate W2N framework to an iterative refinement process including several **localization adaptation modules** and **semi-supervised learning modules**. In the localization adaptation module, we initialize a fully supervised detector training on the noisy dataset generated by WSOD. During the training phase, we generate a proposal outside each noisy box annotation and then store the decoded boxes of their regression results. Meanwhile, the decoded boxes are used to calculate a regularization loss to optimize the detector. After training, we use this detector to generate pseudo ground-truth again to reduce the proportion of bounding box located at discriminative part and then step into the semi-supervised learning module. And in the semi-supervised learning module, we first split the dataset with pseudo ground-truths into labeled set and a unlabeled set by the hybrid-level dataset split method. And then a semi-supervised object detection framework is performed to train a detector on these two sets. Finally, we execute these two modules iteratively and construct an iterative training framework for better detection performance with only image-level annotations.

Extensive experiments and ablation studies have been conducted to evaluate the effectiveness our proposed method. The experimental results demonstrate that our W2N framework brings huge improvement for all baselines on different benchmark datasets. In conclusion, the contributions of this paper are summarized as follows:

1) We propose a new multi-phase WSOD paradigm, which formulates the multi-phase weakly supervised object detection problem as a noisy-label object detection problem to reduce the negative effect from low-quality pseudo ground-truths.
2) To tackle the noisy-label training problem, we proposed an iterative learning framework including localization adaptation module and semi-supervised learning module, which improves the quality of pseudo ground-truths and the performance of detector.
3) Experimental results on different benchmark datasets show that our proposed method bring a huge improvement for all WSOD baseline and achieve state-of-the-art performance on WSOD tasks.

2 Related Work

2.1 Weakly Supervised Object Detection

Existing WSOD methods [2,31,30,27] are usually based on multiple instance learning (MIL) [7], which formulate this task as a proposal classification problem. Nevertheless, most of the WSOD algorithms tend to recognize the discriminative parts of some objects and optimizing into local-minima, which promote the proposals of several approaches [4,11,25]. Recently, some works [8,42,3] have

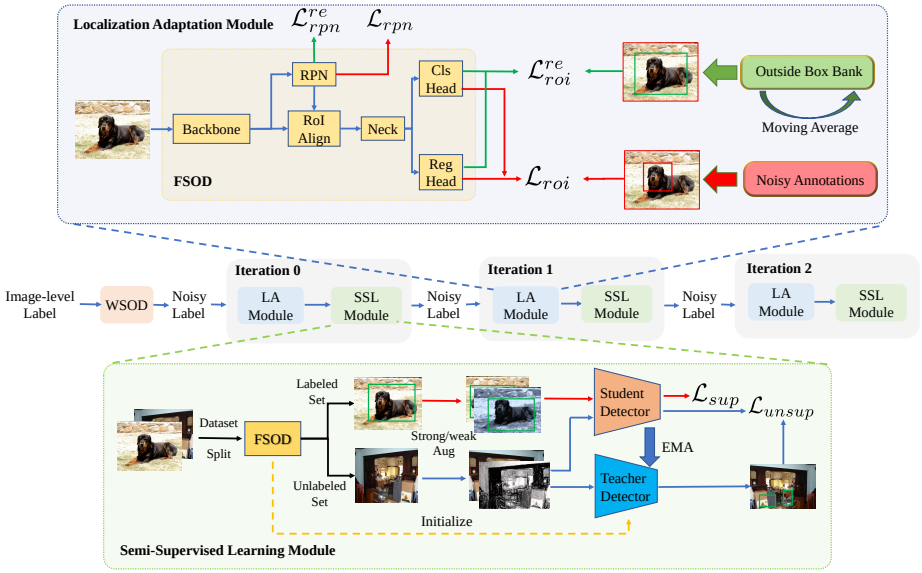

Fig. 2: The illustration of our Weak-to-Noisy (W2N) method, which executes localization adaptation modules (LA module) and semi-supervised learning modules (SSL module) iteratively to generate more accurate pseudo labels and supervise a better object detector. Specifically, the localization adaptation module focus on handling bounding boxes of discriminative parts in $\mathbb{X}_p$ to enlarge the corresponding bounding box and cover more parts of the object, and the semi-supervised learning module leverages the pseudo ground-truth of $\mathbb{X}_p$ with higher detection precision to enhance the final detection performance.

leveraged transfer learning paradigm with an external fully-annotated source dataset to further improve the detection performance of WSOD. In addition, some work managed to convert weak supervision into other paradigms. For example, W2F [41] combined the weakly-supervised detector and the fully-supervised detector by our pseudo ground-truth mining algorithm. SoS [29] harness all potential supervisory signals in WSOD and split the dataset into labeled and unlabeled images to execute a SSOD framework. To the best of our knowledge, we are the first to formulate the weakly supervised object detection problem as a noisy-label object detection problem. In addition, we explore the noise characteristic of every instance-level annotation and design two learning modules to enhance their accuracy, which is not explored in previous works.

2.2 Learning with Noisy Labels

Some work are engaged in exploring how to train an image classifier with noisy labels. To address this problem, DivideMix [16] used two networks to perform

sample selection via a two-component mixture model. Pleiss *et al.*[23] introduced the *Area Under the Margin* statistic which measures the average difference between the logit of a sample's assigned class and its highest non-assigned to separate correctly-labeled data from mislabeled data. Liu *et al.*[20] found that model learns to predict the true labels during the early learning stage but eventually memorizes the wrong labels, which inspires them to leverage the early output of the model. We absorb the inspiration of these work and adapt them to noisy label object detection framework.

2.3 Semi Supervised Object Detection

Semi-supervised learning aims to training networks with both a few of labeled and amount of unlabeled data. In this setting, Jeong *et al.*[12] proposed a consistency-based method, which enforces the predictions of an input image and its flipped version to be consistent. STAC [28] proposes to use a weak data augmentation for model training and a strong data augmentation is used for performing pseudo-label. Liu *et al.*[22] proposed a simple yet effective method, Unbiased Teacher, to address the pseudolabeling bias issue caused by class-imbalance existing in ground-truth labels and the overfitting issue caused by the scarcity of labeled data. Xu.*et al.*[36] proposed a soft teacher mechanism as well as a box jittering approach to improve the overall detection performance with semi-supervised manner.

3 Proposed Method

Definition. Let $\mathbb{X} = \{(\mathbf{I}, \mathbb{P}, \mathbf{y})\}$ denotes the weakly annotated dataset including C individual object categories, where $\mathbf{I}$ means the input image, $\mathbb{P}$ means the set of proposals w.r.t. $\mathbf{I}$, and $\mathbf{y} = [y_1, y_2, \ldots, y_C]^T$ is the image classification label. WSOD targets at learning an object detector g with only image-level supervision.

3.1 Overview

With given dataset $\mathbb{X}$, we first train a weakly supervised object detector g following previous state-of-the-art methods [25,11,8,31] and then adopt the multi-phase training strategy [41] to generate pseudo ground-truth (PGT) on the training images. Now we obtain a new dataset with supervised signal: $\mathbb{X}_p = \{(\mathbf{I}, \{\mathbf{S}\})\}$, $\mathbf{S} = (\mathbf{b}, c)$, where $\mathbf{b} = [x, y, w, h]$ denotes the instance-level bounding box by its center coordinate (x, y), width w, height h, and c denotes the category of this box. We propose $\mathbb{X}_p$ can be regarded as a noisy annotation due to the low accuracy in terms of classification or localization, and the WSOD task can be converted to an object detection task with noisy annotations. To train an object detector on such noisy dataset, we propose a novel training framework namely Weakly-to-Noisy (W2N), which executes localization adaptation modules and semi-supervised learning modules iteratively to generate more accurate pseudo labels and supervise a better object detector. The overall pipeline of

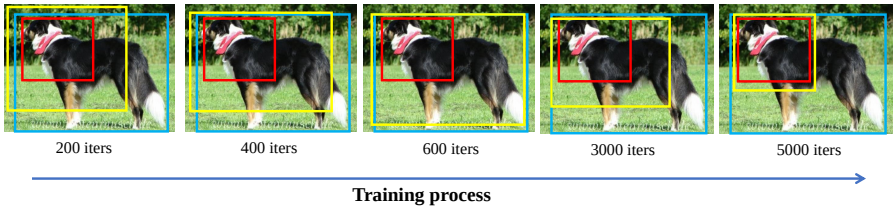

Fig. 3: An example of regression results of a proposals outside the discriminative part pseudo ground-truth during training. Blue box indicates the real ground-truth, red box indicates the discriminative part pseudo ground-truth and the yellow box indicates the outer box of the red one. Yellow box is regressed to the blue box at early stage of training process, but finally overfits to the red box.

W2N is shown as Fig.2. Specifically, the localization adaptation module focus on handling discriminative parts bounding box in $\mathbb{X}_p$ to enlarge the corresponding bounding box and cover more parts of the object, and the semi-supervised learning module leverages the high-quality part of the pseudo ground-truths in $\mathbb{X}_p$ to enhance the final detection performance of object detector.

3.2 Noisy Label Generation

Due to the lack of instance-level supervision during the training procedure of WSOD, the prediction results from the pretrained WSOD network g is not accurate enough [2,31,41], *e.g.*, the wrong prediction in Fig. 4, mislabel or low location accuracy. Following [41,29], we treat the pretrained object detector g as a generator of noisy labels to generate the pseudo ground-truths. We select three WSOD baseline methods to play the role of generators: *OICR+REG* [31], *CASD*[11], and *LBBA*[8]. After training on $\mathbb{X}$, the weakly-supervised detector g inference on training Image $\mathbf{I}$ and we filter the original predictions, convert it to pseudo ground-truth and obtain $\mathbb{X}_p$ according to the Pseudo Ground-Truth Excavation method proposed by W2F[41].

3.3 Learning Detector with Noisy Annotations

After generating the noisy labels, we feed the labels into the W2N training framework to supervise better object detector progressively. Following [16,20], we propose an training framework W2N, which iterates between localization adaptation module and semi-supervised learning module for several steps. The following subsections will illustrate these two modules in details.

Localization Adaptation Module. In semi-supervised learning module which will be mentioned below, the quality of labeled set will effect the performance of the detector [29]. The more accurate label in labeled set, the higher performance

Fig. 4: An example of noisy label. Notice that the orange box has precise bounding box but mislabeled to bicycle (the ground-truth is motorbike), while the category of red box is correct but its bounding box is incorrect.

the model achieve. However, we argue that the dataset split can not recognize and filter the discriminative-part noisy labels among several categories (e.g., like the "person" prediction box Fig 4. The main reason is that too many discriminative-part noisy labels appear in the $\mathbb{X}_p$ such that network tends to overfit them easily during training and then obtain low detection precision.

To deal with this problem, we revisit the characteristic of discriminative-part noisy labels and dig out such regular pattern, which is shown in Fig. 3. First, the discriminative-part noisy labels are usually inside the corresponding real ground-truths. Second, if we use the $\mathbb{X}_p$ to train a supervised object detector f, the outer proposals of the discriminative part noisy labels will regress toward the real ground-truth during the early stage during of training phase. But as training continues, it tends to overfit toward the discriminative part noisy labels again. Based on this observation, we refer to the method of using early output in noisy-label image classification task and design a regularization loss to handle the "discriminative part problem".

As mentioned above, with regard to a discriminative part noisy labels, their corresponding outer proposals will regress toward a more accurate location at early stage learning phase. Therefore we store these proposals as the extra supervision to optimize the fully supervised detector f. Specifically, given a pseudo ground-truth box $\mathbf{b} = [x, y, w, h]$ at iteration t during training phase, we randomly generate a outer box extending from $\mathbf{b}$ it by random sampling the transformation δ^t:

$$\begin{aligned} \delta_x^t, \delta_y^t &\sim \mathcal{U}(-\alpha, \alpha) \\ \delta_w^t, \delta_h^t &\sim \mathcal{U}(\sqrt{3}, 2) \end{aligned} \tag{1}$$

$\mathcal{U}(-\alpha, \alpha)$ denotes an uniform distribution in the range $[-\alpha, \alpha]$. Then a random outer box $\tilde{\mathbf{b}}^{\mathbf{t}} = [\tilde{x}^t, \tilde{y}^t, \tilde{w}^t, \tilde{h}^t]$ is obtained by:

$$[\tilde{x}^t, \tilde{y}^t, \tilde{w}^t, \tilde{h}^t] = [x + \delta_x^t \cdot w, y + \delta_y^t \cdot h, w \cdot \delta_w^t, h \cdot \delta_h^t]. \tag{2}$$

The outer boxes $\tilde{\mathbf{b}}^{\mathbf{t}}$ are fed into the object detector and then obtain the decode boxes $\hat{\mathbf{b}}^{\mathbf{t}}$. To measure the quality of $\hat{\mathbf{b}}^{\mathbf{t}}$, we only select the boxes whose prediction

scores are higher than a threshold τ_{score} while the IoU with corresponding $\mathbf{b}$ are lower than the label assigning threshold τ_{assign} (e.g., 0.5). Finally, to obtain more precision outer boxes, we adopt the moving average strategy to synthesize all $\hat{\mathbf{b}}$ before iteration t and obtain the extra supervision for regularization, shown as Eqn. (3):

$$\hat{\mathbf{b}}_{re}^{t} = \beta \hat{\mathbf{b}}_{re}^{t-1} + (1-\beta)\hat{\mathbf{b}^{\mathbf{t}}}, \tag{3}$$

where β is the moving average value of bounding box. Then we use $\{(\mathbf{b}, c)\}$ and $\{(\hat{\mathbf{b}}_{re}^{t}, c)\}$ as the supervision signal to optimize detector f, and calculate loss function $\mathcal{L}_{rpn}$, $\mathcal{L}_{roi}$, $\mathcal{L}_{rpn}^{re}$ and $\mathcal{L}_{roi}^{re}$, where $\mathcal{L}_{rpn}$ and $\mathcal{L}_{roi}$ indicate the loss supervised with noisy labels $\{(\mathbf{b}, c)\}$ of RPN and RoI head while $\mathcal{L}_{rpn}^{re}$ and $\mathcal{L}_{roi}^{re}$ is calculated with extra supervision $\{(\hat{\mathbf{b}}_{re}^{t}, c)\}$ as regularization terms. Each of them is the combination of Smooth L1 Loss(regression loss) and Cross-Entropy Loss(classification loss), which are the same formulation as [24]. The whole loss function $\mathcal{L}_{fsod}$ for optimization f is shown as Eqn. (4):

$$\mathcal{L}_{fsod} = \mathcal{L}_{rpn} + \mathcal{L}_{roi} + \lambda_{re}(\mathcal{L}_{rpn}^{re} + \mathcal{L}_{roi}^{re}) \tag{4}$$

where λ_{re} indicates the regularization weight.

After the process above, we use the well-trained detector f to re-generate the pseudo ground-truths on the training set, which can reduce the proportion of low-quality pseudo ground-truths and improve the performance of the next semi-supervised learning module.

Semi-Supervised Learning Module. In this module, we design a hybrid-level dataset split algorithm as well as a pseudo-label based semi-supervised training algorithm.

Dataset split method is crucial for turning noisy-label learning into semi-supervised approach. A basic solution is that spliting the whole dataset according to the training loss of each image. The training data with small loss is regarded as the sample from labeled set, vise versa. SoS [29] proposed the "image-level split method", which accumulated the losses from the RPN module and that from the detection head and then obtained the image-level split loss function. Given image $\mathbf{I}$, the image-level split loss $\mathcal{L}_{split}(\mathbf{I})$ is defined as Eqn. (5):

$$\mathcal{L}_{split}(\mathbf{I}) = \underset{i}{\text{avg}}(\mathcal{L}_{split}^{rpn}(R_i, t_i)) + \underset{j}{\text{avg}}(\mathcal{L}_{split}^{roi}(R_j, t_j)). \tag{5}$$

And the $\mathcal{L}_{split}^{rpn}$ and $\mathcal{L}_{split}^{roi}$ are shown as Eqn. (6) and (7):

$$\mathcal{L}_{split}^{rpn}(R_i, t_i) = \mathcal{L}_{rpn}^{cls}(R_i, t_i) + \mathcal{L}_{rpn}^{reg}(R_i, t_i), \tag{6}$$

$$\mathcal{L}_{split}^{roi}(R_j, t_j) = \mathcal{L}_{roi}^{cls}(R_j, t_j) + \mathcal{L}_{roi}^{reg}(R_j, t_j), \tag{7}$$

where R_i is the i-th foreground RoI, t_i indicates the assigned target label of R_i, $\mathcal{L}_{rpn}$ and $\mathcal{L}_{roi}$ are RPN and RoI head losses, and cls and reg stand for

classification task and box regression task, respectively. $\mathcal{L}_{*}^{cls}$ is Cross-Entropy Loss and $\mathcal{L}_{*}^{reg}$ is Smooth L1 Loss. And the avg$(\cdot)$ means the mean average operation. Then, we rank all instances with their $\mathcal{L}_{split}(\mathbf{I})$ by the ascending order, keeping the number of p percent of image annotations with small loss value to be the labeled set. However, we find that a training image may contain multiple instance labels, and the accurate labels and noisy labels often appear at the same time. Therefore, we proposed the second split method namely "instance-level split method", in which every instance is be seen to the smallest split unit. And the aggregated loss in Eqn. (5) will be modified to Eqn. (8):

$$\mathcal{L}_{split}(\mathbf{S}) = \underset{i}{\text{avg}}(\mathcal{L}_{split}^{rpn}(R_i, t_i)) + \underset{j}{\text{avg}}(\mathcal{L}_{split}^{roi}(R_j, t_j)), \tag{8}$$

where $\mathbf{S}$ indicates to one instance label and avg$(\cdot)$ means the mean average operation. Then we rank all instances according to the $\mathcal{L}_{split}(\mathbf{S})$ by the ascending order, and then keep the top p percent of the instance labels with small loss value to be the labeled set $\mathbb{X}_l = \{(\mathbf{I}_l, \{\mathbf{S}_l\})\}$, and the other instances are keeping unlabeled.

In SoS [29], the labeled set are used for supervising the training for classification and regression sub-tasks. However, we can not make sure that each pseudo label is correct in terms of both classification and localization. As shown in Fig. 4, a box with high location information may be mislabeled of category while a box with correct category may cover part of an object. From this perspective, we introduce two tags $\lambda_{cls}, \lambda_{reg}$ for one instance label indicating their confidence for two sub-task respectively. The final formulation of labeled set is modified to $\mathbb{X}_l = \{(\mathbf{I}_l, \{(\mathbf{S}_l, \lambda_{cls}, \lambda_{reg})\})\}$, where $\lambda_{cls}, \lambda_{reg} \in \{0, 1\}, \lambda_{cls} + \lambda_{reg} \neq 0$. $\lambda_{cls} = 1$ means the category label of this instance is correct, while $\lambda_{cls} = 0$ means not, similar meaning for λ_{reg}. To decide the value of $\lambda_{cls}, \lambda_{reg}$, we propose "two tasks instance-level split" method, which is shown as Eqn. (9):

$$\begin{aligned} \mathcal{L}_{split}^{cls}(\mathbf{S}) &= \underset{i}{\text{avg}}(\mathcal{L}_{rpn}^{cls}(R_i, t_i)) + \underset{j}{\text{avg}}(\mathcal{L}_{roi}^{cls}(R_j, t_j)), \\ \mathcal{L}_{split}^{reg}(\mathbf{S}) &= \underset{i}{\text{avg}}(\mathcal{L}_{rpn}^{reg}(R_i, t_i)) + \underset{j}{\text{avg}}(\mathcal{L}_{roi}^{reg}(R_j, t_j)), \end{aligned} \tag{9}$$

where $\mathcal{L}_{split}^{cls}(\mathbf{S})$ only accumulates the classification loss for each foreground proposal while $\mathcal{L}_{split}^{reg}(\mathbf{S})$ only accumulates the regressions loss for each foreground proposal. Then, we rank the instance according to $\mathcal{L}_{split}^{cls}(\mathbf{S})$ and $\mathcal{L}_{split}^{reg}(\mathbf{S})$ by the ascending order, respectively. Finally, we set $\lambda_{cls} = 1$ for the top p percent of the instances in terms of $\mathcal{L}_{split}^{cls}(\mathbf{S})$ and set $\lambda_{reg} = 1$ for the top p percent of the instances in terms of $\mathcal{L}_{split}^{reg}(\mathbf{S})$. In Sec. 4, we will discuss the effect of three data split proposed above.

After spliting the noisy dataset, we introduce a novel semi-supervised object detection method for weakly-to-noisy label training. The difference between [22] and our semi-supervised detection method is two-fold. First, we use labeled set $\mathbb{X}_l$ as labeled set to optimize model with the supervised loss $\mathcal{L}_{sup}$. Combining with our *two tasks instance-level split* method, we modify the origin supervised

loss function with adding the value of $(\lambda_{cls}, \lambda_{reg})$. Specifically, $\mathcal{L}_{sup}$ is shown as Eqn. (10):

$$\begin{aligned}\mathcal{L}_{sup}(\mathrm{I}) = &\operatorname*{avg}_i(\lambda_{cls}^{t_i}\mathcal{L}_{rpn}^{cls}(R_i, t_i) + \lambda_{reg}^{t_i}\mathcal{L}_{rpn}^{reg}(R_i, t_i)) \\ &+ \operatorname*{avg}_j(\lambda_{cls}^{t_j}\mathcal{L}_{roi}^{cls}(R_j, t_j) + \lambda_{reg}^{t_j}\mathcal{L}_{roi}^{reg}(R_j, t_j)) \\ &+ \operatorname*{avg}_k(\mathcal{L}_{bg}(R_k)),\end{aligned} \tag{10}$$

where $\mathcal{L}_{bg}(R_i)$ indicates the background loss of corresponding proposals. Particularly, only the target label of which $\lambda_{cls}^{t_i} = 1$ ($\lambda_{reg}^{t_i} = 1$) can contribute to $\mathcal{L}_{sup}$ in classification (regression) task. The loss function used on the labeled set is shown as Eqn. (11):

$$\mathcal{L}_{sup} = \frac{1}{N_l}\sum_i \mathcal{L}_{sup}(\mathrm{I}_i), \tag{11}$$

where N_l is the number of image in $\mathbb{X}_l$. Second, the regression loss of the unlabeled data are not adopted in the whole training process of [22]. In our method we adopt the box jittering strategy proposed by [36] and add the regression loss of the unlabeled data in origin $\mathcal{L}_{unsup}$ [22]. Finally, the whole loss function of SSOD module is shown as Eqn. (12):

$$\mathcal{L}_{ssod} = \mathcal{L}_{sup} + \lambda_u\mathcal{L}_{unsup}, \tag{12}$$

where λ_u is the weight of $\mathcal{L}_{unsup}$.

Iterative Training Framework Finally, we propose the two-phase iterative training framework based on these two modules. The whole training process of our framework is given in Algorithm 1, which is summarized as follows. Specifically, the first phase is the conventional weakly-supervised object detection pre-training module, we train a WSOD network g and then generated the pseudo ground-truths for each training image in the training dataset $\mathbb{X}_p^0$. The second phase is our proposed weakly-to-noisy training framework. Given the pseudo ground-truths, we first execute the localization adaptation module to initialize a fully-supervised detector f_t and then refine $\mathbb{X}_p^t$ to reduce the proportion of the discriminative part. Then we excute the two tasks instance-level split method and split the whole training set $\mathbb{X}_p^t$ into labeled set and unlabeled set. With the splitted training sets, we execute the semi-supervised object detection module to supervise a better object detector $f_t^{'}$. Generally, we use $f_t^{'}$ to update the $\mathbb{X}_p^t$ to $\mathbb{X}_p^{t+1}$ and then perform these two modules iteratively for T times. And finally, the last object detector $f_T^{'}$ with corresponding parameters θ_f^T is saved for usage.

4 Experiments

4.1 Experiment Settings

Datasets. Following [25,32,42], we evaluate our method on four benchmarks: PASCAL VOC 2007, PASCAL VOC 2012 [9], MS-COCO [19], and ILSVRC

Algorithm 1 Weak Supervision to Noisy Supervision for Object Detection

Input: Iteration number T, weakly annotated dataset $\mathbb{X}$;
Output: An updated detector $f_T^{'}$;
1: Train the weakly supervised detector g on $\mathbb{X}$;
2: Obtain the noisy annotations dataset $\mathbb{X}_p^0$ by pretrained weakly supervised detector g;
3: **for** $t = 0...T-1$ **do**
4: **Localization Adaptation module:**
5: Initialize an object detector f_t on $\mathbb{X}_p^t$;
6: Refine $\mathbb{X}_p^t$ by f_t;
7: **Semi-Supervised Learning module:**
8: Split $\mathbb{X}_p^t$ into labeled set and unlabeled set by f_t;
9: Execute the semi-supervised object detection approach to optimize f_t to $f_t^{'}$;
10: Update the $\mathbb{X}_p^t$ to $\mathbb{X}_p^{t+1}$ by $f_t^{'}$;

2013 [6] detection dataset. **Evaluation Metrics.** We use mean average precision (mAP) to evaluate the detection performance over categories, and CorLoc to measure the localization accuracy.

4.2 Comparison with State-of-the-arts

We *state the implementation details in the suppl.* And here we compare our method with several state-of-the-art WSOD approaches in terms of mAP and CorLoc on PASCAL VOC 2007 [9] reported by Table 1 and Table 2. Our all results are obtained with single-scale testing approch. Based on these results, we obtain the following observations: First, our W2N framework outperforms all WSOD baselines in terms of both mAP and CorLoc. Specifically, on PASCAL VOC 2007 dataset, it outperforms OICR+REG by 8.7% mAP and 3.8% CorLoc, outperforms CASD by 11.4% mAP and 12.6% CorLoc, and outperforms LBBA by 9.5% mAP and 10.8% CorLoc. Performance on PASCAL VOC 2012 also demonstrates favorable performance improvement.

Second, our W2N outperforms all of the state-of-the-art WSOD methods as well as transfer learning based methods. Specifically, CASD+W2N achieves 65.4% mAP on PASCAL VOC 2007 test set, which outperforms CASD by 8.6% mAP and outperforms CaT_5 by 1.9% mAP. Moreover, LBBA+W2N obtains 68.6% mAP and 83.4% CorLoc, which achieves a new state-of-the-arts for WSOD problem and bridges the performance gap with fully supervised methods (Faster R-CNN)[24]. In the supplementary we will show more results on other datasets and give analyze for comparison between [29] and ours.

4.3 Ablation Study

In this section, we discuss the effect of key components of W2N on PASCAL VOC 2007 dataset [9].
Effect of two modules.

Table 1: Comparison of our method on PASCAL VOC 2007 test set to state-of-the-art WSOD methods in terms of mAP (%), where $^+$ means the results with multi-scale testing.

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	AP
Pure WSOD:																					
WSDDN [2]	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
OICR$^+$ [31]	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$^+$ [30]	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
Yang *et al.*$^+$ [38]	57.6	70.8	50.7	28.3	27.2	72.5	69.1	65.0	26.9	64.5	47.4	47.7	53.5	66.9	13.7	29.3	56.0	54.9	63.4	65.2	51.5
C-MIDN$^+$ [37]	53.3	71.5	49.8	26.1	20.3	70.3	69.9	68.3	28.7	65.3	45.1	64.6	58.0	71.2	20.0	27.5	54.9	54.9	69.4	63.5	52.6
Arun *et al.* [1]	66.7	69.5	52.8	31.4	24.7	74.5	74.1	67.3	14.6	53.0	46.1	52.9	69.9	70.8	18.5	28.4	54.6	60.7	67.1	60.4	52.9
WSOD2$^+$ [40]	65.1	64.8	57.2	39.2	24.3	69.8	66.2	61.0	29.8	64.6	42.5	60.1	71.2	70.7	21.9	28.1	58.6	59.7	52.2	64.8	53.6
GradingNet-C-MIL [13]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54.3
MIST-Full [25]	68.8	77.7	57.0	27.7	28.9	69.1	74.5	67.0	32.1	73.2	48.1	45.2	54.4	73.7	35.0	29.3	64.1	53.8	65.3	65.2	54.9
IM-CFB$^+$ [39]	63.3	77.5	48.3	36.0	32.6	70.8	71.9	73.1	29.1	68.7	47.1	69.4	56.6	70.9	22.8	24.8	56.0	59.8	73.2	64.6	55.8
CASD [11]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	56.8
SoS [29]	72.9	79.4	59.6	20.4	49.8	81.2	82.9	84.0	31.5	76.6	57.4	60.7	74.7	75.1	33.0	34.3	66.3	61.1	80.6	71.8	62.7
SoS$^+$ [29]	77.9	81.2	58.9	26.7	54.3	82.5	84.0	83.5	36.3	76.5	57.5	58.4	78.5	78.6	33.8	37.4	64.0	63.4	81.5	74.0	64.4
OICR+REG (reproduce)	54.0	61.9	43.9	22.6	31.7	73.8	65.1	60.6	14.4	68.0	17.0	48.8	58.3	69.9	12.8	22.0	53.9	53.6	69.7	60.4	48.3
CASD (reproduce)	68.8	67.2	53.9	38.2	21.5	70.4	69.7	68.9	23.6	66.3	48.8	62.3	56.4	70.6	17.3	24.9	55.9	58.9	66.0	69.1	54.0
OICR+REG+W2N (Ours)	71.0	74.2	60.8	28.8	44.6	78.0	72.6	80.3	16.7	74.3	24.3	58.2	64.6	75.1	13.3	29.9	60.3	65.3	80.1	67.6	**57.0(+8.7)**
CASD+W2N (Ours)	74.0	81.7	71.2	48.9	51.0	78.6	82.3	83.5	29.1	76.9	51.5	82.1	76.9	79.1	28.5	34.3	65.0	64.2	75.2	74.8	**65.4(+11.4)**
WSOD with transfer learning:																					
MSD-Ens$^+$ [18]	70.5	69.2	53.3	43.7	25.4	68.9	68.7	56.9	18.4	64.2	15.3	72.0	74.4	65.2	15.4	25.1	53.6	54.4	45.6	61.4	51.1
OICR+UBBR [14]	59.7	44.8	54.0	36.1	29.3	72.1	67.4	70.7	23.5	63.8	31.5	61.5	63.7	61.9	37.9	15.4	55.1	57.4	69.9	63.6	52.0
LBBA$^+$ [8]	70.3	72.3	48.7	38.7	30.4	74.3	76.6	69.1	33.4	68.2	50.5	67.0	49.0	73.6	24.5	27.4	63.1	58.9	66.0	69.2	56.6
Zhong *et al.* (R50-C4)$^+$ [42]	64.8	50.7	65.5	45.3	46.4	75.7	74.0	80.1	31.3	77.0	26.2	79.3	74.8	66.5	57.9	11.5	68.2	59.0	74.7	65.5	59.7
TraMaS$^+$ [21]	68.6	61.1	69.6	48.1	49.9	76.3	77.8	80.9	34.9	77.0	31.1	80.9	78.5	66.3	64.0	19.1	69.1	62.3	74.4	69.1	62.9
CaT$_5$ [3]	74.0	70.7	60.0	31.1	50.0	75.9	82.0	70.7	32.8	74.3	69.5	70.2	69.5	77.0	37.5	45.8	67.0	61.1	72.4	68.0	63.0
LBBA (reproduce)	70.2	75.5	49.2	41.9	30.5	80.5	78.2	72.8	36.4	73.8	52.3	67.0	46.4	76.2	34.6	29.4	67.9	66.6	68.3	74.1	59.1
LBBA+W2N (Ours)	71.8	83.0	69.9	50.3	54.5	79.0	83.9	83.9	39.4	79.2	52.9	82.2	83.6	79.2	62.6	32.7	68.5	66.1	75.8	74.5	**68.6(+9.5)**
Upper bounds:																					
Faster R-CNN (Res50+FPN) [24]	82.8	84.2	75.2	62.4	67.0	81.4	87.1	82.6	57.3	82.5	64.9	83.0	84.0	82.7	83.7	54.0	76.1	73.4	81.8	76.1	**76.1**

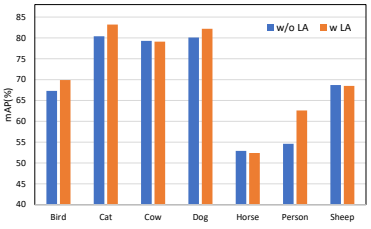

Fig. 5: Effect of location adaption module on animal categories and person category with LBBA+W2N.

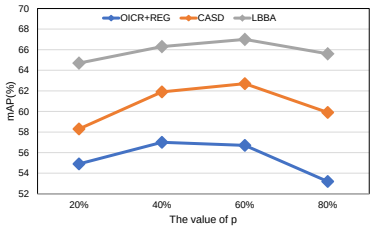

Fig. 6: Effect of different size of the labeled set on VOC 2007 for different WSOD+W2Ns

Table 3 shows the ablation study of each module on LBBA baseline. Simply re-training Faster R-CNN(FRCNN*) with pseudo GT only brings 0.3% mAP gain. By introducing localization adaption and semi-supervised learning separately, these improvements respectively outperform the baseline by 1.2% and 7.0% in terms of mAP. Specifically, as illustrated in Fig. 5, W2N+LBBA with location adaption module improves the detection performance of categories which suffer from the discriminative part problem, especially for person category. Furthermore, our full method combining these two modules can further improve the detection performance to 67.0% mAP. More ablation study about effect of two modules can be found in the suppl..

Effect of Iterative Training. Generally, more training iterations means better predictions. Thus we analyze the effect of training iteration T. Table. 4 shows the performance of W2N with different iteration numbers T using three differ-

Table 2: Comparison of our method on PASCAL VOC 2007 trainval set to state-of-the-art WSOD methods in terms of CorLoc (%), where $^+$ means the results with multi-scale testing.

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	CorLoc
Pure WSOD:																					
WSDDN [2]	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
OICR$^+$ [31]	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$^+$ [30]	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
Li$^+$ [17]	85.0	83.9	58.9	59.6	43.1	79.7	85.2	77.9	31.3	78.1	50.6	75.6	76.2	88.4	49.7	56.4	73.2	62.6	77.2	79.9	68.6
C-MIL$^+$ [34]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.0
Yang *et al.*$^+$ [38]	80.0	83.9	74.2	53.2	48.5	82.7	86.2	69.5	39.3	82.9	53.6	61.4	72.4	91.2	22.4	57.5	83.5	64.8	75.7	77.1	68.0
MIST (Full)$^+$ [25]	87.5	82.4	76.0	58.0	44.7	82.2	87.5	71.2	49.1	81.5	51.7	53.3	71.4	92.8	38.2	52.8	79.4	61.0	78.3	76.0	68.8
WSOD2$^+$ [40]	87.1	80.0	74.8	60.1	36.6	79.2	83.8	70.6	43.5	88.4	46.0	74.7	87.4	90.8	44.2	52.4	81.4	61.8	67.7	79.9	69.5
Arun *et al.*[1]	88.6	86.3	71.8	53.4	51.2	87.6	89.0	65.3	33.2	86.6	58.8	65.9	87.7	93.3	30.9	58.9	83.4	67.8	78.7	80.2	70.9
GradingNet-C-MIL [13]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	72.1
IM-CFB$^+$ [39]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	72.2
OICR+REG (reproduce)	91.6	78.3	62.6	46.0	44.8	86.4	87.7	80.3	34.4	87.1	30.1	69.4	81.1	90.8	31.3	44.8	76.0	76.1	83.1	60.5	67.4
CASD (reproduce)	68.8	67.2	53.9	38.2	21.5	70.4	69.7	68.9	23.6	66.3	48.8	62.3	56.4	70.6	17.3	24.9	55.9	58.9	66.0	69.1	68.5
OICR+REG+W2N (Ours)	87.4	86.0	69.7	50.8	59.8	89.8	88.4	86.9	37.5	86.5	26.0	69.8	84.0	95.1	31.6	57.6	78.12	75.6	85.8	77.3	**71.2(+3.8)**
CASD+W2N (Ours)	92.0	90.5	82.4	71.3	73.0	85.5	94.7	89.0	46.3	89.4	63.5	87.9	92.7	96.7	47.1	70.2	84.4	75.1	82.4	87.5	**80.1(+12.6)**
WSOD with transfer learning:																					
OICR+UBBR [14]	47.9	18.9	63.1	39.7	10.2	62.3	69.3	61.0	27.0	79.0	24.5	67.9	79.1	49.7	28.6	12.8	79.4	40.6	61.6	28.4	47.6
WSLAT-Ens [26]	78.6	63.4	66.4	56.4	19.7	82.3	74.8	69.1	22.5	72.3	31.0	63.0	74.9	78.4	48.6	29.4	64.6	36.2	75.9	69.5	58.8
MSD-Ens$^+$ [18]	89.2	75.7	75.1	66.5	58.8	78.2	88.9	66.9	28.2	86.3	29.7	83.5	83.3	92.8	23.7	40.3	85.6	48.9	70.3	68.1	66.8
Zhong *et al.* (R50-C4)$^+$ [42]	87.5	64.7	87.4	69.7	67.9	86.3	88.8	88.1	44.4	93.8	31.9	89.1	92.9	86.3	71.5	22.7	94.8	56.5	88.2	76.3	74.4
LBBA$^+$ [8]	93.3	90.6	71.8	69.2	59.5	90.9	94.4	78.5	55.4	96.6	51.0	82.3	72.5	93.2	48.5	52.8	100.0	66.7	78.3	87.5	76.7
TraMaS$^+$ [21]	90.6	67.4	89.7	70.5	72.8	86.6	91.7	89.8	51.0	96.1	34.0	93.7	94.8	90.3	73.0	26.5	95.2	68.2	89.8	83.1	77.7
CaT$_5$ [3]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	80.3
LBBA (reproduce)	86.9	84.5	74.6	65.6	55.1	85.4	86.8	84.4	42.5	88.0	45.0	83.3	82.3	88.6	47.6	49.1	88.3	50.8	81.1	84.3	72.7
LBBA+W2N (Ours)	89.5	93.4	83.9	70.2	73.4	87.1	94.5	92.0	58.9	95.7	64.0	91.0	94.8	93.5	80.7	64.1	91.7	78.2	84.3	89.1	**83.5(+10.8)**
Upper bounds:																					
Faster R-CNN (Res50+FPN)[24]	91.7	93.7	92.6	75.0	84.0	95.4	95.3	93.2	76.5	94.5	86.9	92.3	96.0	93.2	93.0	76.8	94.9	89.2	85.7	90.4	**89.5**

Table 3: Effect of two modules on VOC 2007.

WSOD	FRCNN*	LA	SSL	ITER	mAP
LBBA					59.1
	✓				59.4
		✓			60.3
			✓		66.1
		✓	✓		67.0
		✓	✓	✓	**68.6**

Table 4: The mAP results of our W2N with different iteration times T on Pascal VOC 2007 dataset.

Methods	0	1	2	3	4
OICR+REG+W2N	56.8	**57.0**	56.8	56.8	56.9
CASD+W2N	62.7	64.5	**65.4**	65.4	65.2
LBBA+W2N	67	67.9	**68.6**	68.4	68.4

ent methods, respectively. Generally, as the T increases, the performance first increase and then begin to oscillate near the highest point. And the highest performance for all baseline are outperforming beyond 1.5% mAP than the settings of $T = 0$, which proves that the iterative training strategy is effective for further improving detection performance. In addition, for LBBA and CASD, it reaches the highest performance when $T = 2$; while for OICR+REG, $T = 1$ is the best optimal solution. This result indicates that the iterative training process will converge quickly on relative small T, which reveals the high efficiency of W2N.

Effect of Hybrid-Level Dataset Split. we combined three different WSOD methods with three different split methods and then obtained nine different experiment settings. We conducted experiments on all of the settings at iteration 0 and demonstrate the results in Table. 5. Experimental results prove that the two tasks instance-level split method achieves the best performance among them, higher than the instance-level split method. In addition, both two tasks instance-level split method and instance-level split method outperform the image-level

Table 5: Comparisons of different dataset split methods on VOC 2007.

Methods	image-level	instance-level	two tasks instance-level	mAP
OICR+REG+W2N	✓			55.4
OICR+REG+W2N		✓		56.8
OICR+REG+W2N			✓	**56.8**
CASD+W2N	✓			61.9
CASD+W2N		✓		62.6
CASD+W2N			✓	**62.7**
LBBA+W2N	✓			65.4
LBBA+W2N		✓		66.8
LBBA+W2N			✓	**67.0**

split method more than about 1.5% mAP, which proves that it is more effective and reasonable to treat the instance-level as the smallest division unit.

Proportion of Clean Split p. The proportion of clean split p determines the quality of pseudo labels, therefore here we explore the effect of different p. We deploy varying p to decide the size of labeled set for three different WSOD methods at iteration 0. Fig. 6 shows that for LBBA and CASD, $p = 60\%$ is the best choice, while for OICR+REG, $p = 40\%$ is better. Generally, when p is small, as p increases, the performance of W2Ns improves, while p further increases, the performance of W2Ns begin to drop significantly. This is reasonable that too small leads to a small size of high quality pseudo label in labeled set, which is not conducive to model learning. While too large clean size will involve more noisy labels. Therefore, we propose that a moderate size is beneficial for training.

5 Conclusion

In this paper, we propose a weakly supervised object detection method namely Weakly-supervision to Noisy-supervision (W2N). We treat the pseudo labels generated by the pretrained weakly detector as noisy labels and propose an iterative training procedure, which includes the localization adaptation module and the semi-supervised learning module. The localization adaptation module refines the original pseudo ground-truths to reduce the proportion of low-quality bounding boxes. The semi-supervised learning module split the dataset with pseudo ground-truths into a high-quality labeled set as well as an unlabeled set and supervises the object detector with a well-designed semi-supervised object detection manner with these two datasets. Extensive experiments on different datasets show that our proposed method performs favorably against other state-of-the-art WSOD methods.

Acknowledgement This work was supported in part by the National Key R&D Program of China under Grant No. 2021ZD0112100, and the Major Key Project of PCL under Grant No. PCL2021A12. This work was done when Zitong was an intern at MEGVII Tech.

References

1. Arun, A., Jawahar, C., Kumar, M.P.: Dissimilarity coefficient based weakly supervised object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019). https://doi.org/10.1109/cvpr.2019.00966
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2846–2854 (2016)
3. Cao, T., Du, L., Zhang, X., Chen, S., Zhang, Y., Wang, Y.F.: Cat: Weakly supervised object detection with category transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3070–3079 (October 2021)
4. Chen, Z., Fu, Z., Jiang, R., wu Chen, Y., Hua, X.: Slv: Spatial likelihood voting for weakly supervised object detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12992–13001 (2020)
5. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. IEEE transactions on pattern analysis and machine intelligence **39**(1), 189–203 (2016)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**, 31–71 (1997)
8. Dong, B., Huang, Z., Guo, Y., Wang, Q., Niu, Z., Zuo, W.: Boosting weakly supervised object detection via learning bounding box adjusters. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2876–2885 (2021)
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
10. Girshick, R.: Fast r-cnn. In: International Conference on Computer Vision (ICCV) (2015)
11. Huang, Z., Zou, Y., Bhagavatula, V., Huang, D.: Comprehensive attention self-distillationfor weakly-supervised object detection. In: NeurIPS (2020)
12. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. Advances in neural information processing systems **32**, 10759–10768 (2019)
13. Jia, Q., Wei, S., Ruan, T., Zhao, Y., Zhao, Y.: Gradingnet: Towards providing reliable supervisions for weakly supervised object detection by grading the box candidates. Proceedings of the AAAI Conference on Artificial Intelligence **35**(2), 1682–1690 (May 2021), `https://ojs.aaai.org/index.php/AAAI/article/view/16261`
14. Lee, S., Kwak, S., Cho, M.: Universal bounding box regression and its applications. In: Jawahar, C., Li, H., Mori, G., Schindler, K. (eds.) Computer Vision – ACCV 2018. pp. 373–387. Springer International Publishing, Cham (2019)
15. Li, D., Huang, J.B., Li, Y., Wang, S., Yang, M.H.: Weakly supervised object localization with progressive domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3512–3520 (2016)
16. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. In: International Conference on Learning Representations (2019)

17. Li, X., Kan, M., Shan, S., Chen, X.: Weakly supervised object detection with segmentation collaboration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
18. Li, Y., Zhang, J., Zhang, J., Huang, K.: Mixed supervised object detection with robust objectness transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP** (02 2018). https://doi.org/10.1109/TPAMI.2018.2810288
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
20. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Advances in Neural Information Processing Systems **33** (2020)
21. Liu, Y., Zhang, Z., Niu, L., Chen, J., Zhang, L.: Mixed Supervised Object Detection by Transferring Mask Prior and Semantic Similarity. arXiv e-prints arXiv:2110.14191 (Oct 2021)
22. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: International Conference on Learning Representations (2020)
23. Pleiss, G., Zhang, T., Elenberg, E., Weinberger, K.Q.: Identifying mislabeled data using the area under the margin ranking. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 17044–17056. Curran Associates, Inc. (2020), `https://proceedings.neurips.cc/paper/2020/file/c6102b3727b2a7d8b1bb6981147081ef-Paper.pdf`
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
25. Ren, Z., Yu, Z., Yang, X., Liu, M.Y., Lee, Y.J., Schwing, A.G., Kautz, J.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
26. Rochan, M., Wang, Y.: Weakly supervised localization of novel objects using appearance transfer. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4315–4324 (2015). https://doi.org/10.1109/CVPR.2015.7299060
27. Shen, Y., Ji, R., Wang, Y., Wu, Y., Cao, L.: Cyclic guidance for weakly supervised joint detection and segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
28. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
29. Sui, L., Zhang, C.L., Wu, J.: Salvage of supervision in weakly supervised detection (2021)
30. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.: Pcl: Proposal cluster learning for weakly supervised object detection. IEEE transactions on pattern analysis and machine intelligence **42**(1), 176–191 (2018)
31. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: CVPR (2017)
32. Uijlings, J., Popov, S., Ferrari, V.: Revisiting knowledge transfer for training object class detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1101–1110 (2018)

33. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2), 154–171 (2013)
34. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-mil: Continuation multiple instance learning for weakly supervised object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019)
35. Wan, F., Wei, P., Jiao, J., Han, Z., Ye, Q.: Min-entropy latent model for weakly supervised object detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (Jun 2018). https://doi.org/10.1109/cvpr.2018.00141
36. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. arXiv preprint arXiv:2106.09018 (2021)
37. Yan, G., Liu, B., Guo, N., Ye, X., Wan, F., You, H., Fan, D.: C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9833–9842 (2019). https://doi.org/10.1109/ICCV.2019.00993
38. Yang, K., Li, D., Dou, Y.: Towards precise end-to-end weakly supervised object detection network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8372–8381 (2019)
39. Yin, Y., Deng, J., Zhou, W., Li, H.: Instance mining with class feature banks for weakly supervised object detection. Proceedings of the AAAI Conference on Artificial Intelligence **35**(4), 3190–3198 (May 2021), `https://ojs.aaai.org/index.php/AAAI/article/view/16429`
40. Zeng, Z., Liu, B., Fu, J., Chao, H., Zhang, L.: Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
41. Zhang, Y., Bai, Y., Ding, M., Li, Y., Ghanem, B.: W2f: A weakly-supervised to fully-supervised framework for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 928–936 (2018)
42. Zhong, Y., Wang, J., Peng, J., Zhang, L.: Boosting weakly supervised object detection with progressive knowledge transfer. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 615–631. Springer International Publishing, Cham (2020)