Point-to-Box Network for Accurate Object Detection via Single Point Supervision

Pengfei Chen[1], Xuehui Yu[1], Xumeng Han[1], Najmul Hassan[2], Kai Wang[2], Jiachen Li[2], Jian Zhao[3], Humphrey Shi[2,4], Zhenjun Han[1⋆], and Qixiang Ye[1]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] SHI Lab @ U of Oregon & UIUC, USA
[3] Institute of North Electronic Equipment, Beijing, China
[4] Picsart AI Research (PAIR)
{chenpengfei20, yuxuehui17, hanxumeng19}@mails.ucas.ac.cn
{najmulhassan1628, kk94wang, chrisleesjtu, shihonghui3}@gmail.com
zhaojian90@u.nus.edu, {hanzhj, qxye}@ucas.ac.cn

Abstract. Object detection using single point supervision has received increasing attention over the years. However, the performance gap between point supervised object detection (PSOD) and bounding box supervised detection remains large. In this paper, we attribute such a large performance gap to the failure of generating high-quality proposal bags which are crucial for multiple instance learning (MIL). To address this problem, we introduce a lightweight alternative to the off-the-shelf proposal (OTSP) method and thereby create the Point-to-Box Network (P2BNet), which can construct an inter-objects balanced proposal bag by generating proposals in an anchor-like way. By fully investigating the accurate position information, P2BNet further constructs an instance-level bag, avoiding the mixture of multiple objects. Finally, a coarse-to-fine policy in a cascade fashion is utilized to improve the IoU between proposals and ground-truth (GT). Benefiting from these strategies, P2BNet is able to produce high-quality instance-level bags for object detection. P2BNet improves the mean average precision (AP) by more than 50% relative to the previous best PSOD method on the MS COCO dataset. It also demonstrates the great potential to bridge the performance gap between point supervised and bounding-box supervised detectors. The code will be released at github.com/ucas-vg/P2BNet.

Keywords: Object Detection, Single Point Annotation, Point Supervised Object Detection.

1 Introduction

Object detectors [13,30,29,25,23,4,38,46] trained with accurate bounding box annotations have been well received in academia and industry. However, collecting quality bounding box annotations requires extensive human efforts. To solve

⋆ Corresponding author.

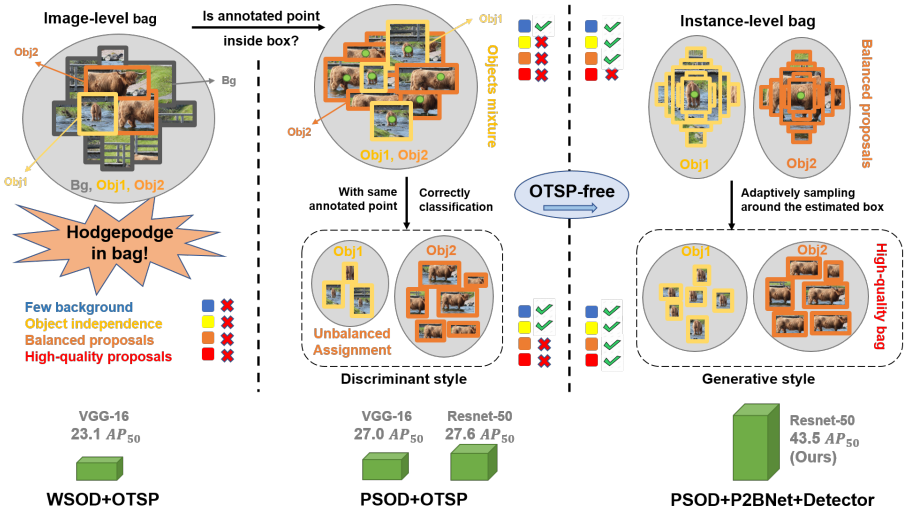
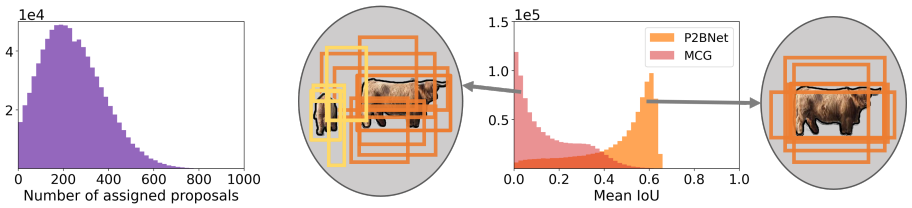


Fig. 1. Based on OTSP methods, the image-level bag in WSOD shows many problems: Too much background, mixture of different objects, unbalanced and low-quality proposals. With point annotation, the previous work UFO^2 filters most background in first stage and splits bags for different objects in refinement. Our P2BNet produces balanced instance-level bags in coarse stage and improves bag quality improves by adaptively sampling proposal boxes around the estimated box of the former stage for better optimization. The performance is the performance in COCO-14. The 27.6 AP_{50} is conducted on UFO^2 with ResNet-50 and our point annotation for a fair comparison.

this problem, weakly supervised object detection [2,39,40,6,41,51,8,49] (WSOD) replace bounding box annotations using low-cost image-level annotations. However, lacking crucial location information and experiencing the difficulty of distinguishing dense objects, WSOD methods perform poorly in complex scenarios. Point supervised object detection (PSOD), on the other hand, can provide distinctive location information about the object and is much cheaper compared with that via bounding box supervision.

Recently, point-based annotations are widely used in many tasks including object detection [28,32] and localization [45,33,37], instance segmentation [7], and action localization [21]. However, the performance gap between point supervised detection methods [28,32] and bounding box supervised detectors remain large. Although it is understandable that location information provided by bounding boxes is richer than the points, we argue that this is not the only reason. We believe most PSOD methods do not utilize the full potential of point-based annotations. Previous works use off-the-shelf proposal (OTSP) methods (*e.g.*, Selective Search [34], MCG [1], and EdgeBox [53]) to obtain proposals for constructing bags. Despite the wide adaptation of these OTSP-based methods in weakly supervised detectors, they suffer from the following problems in Fig. 1: 1) There are too many background proposals in the bags. OTSP methods generate



(a) (b)

Fig. 2. (a) The number of assigned proposal boxes per object produced by MCG (OTSP -based) is unbalanced, which is unfair for training. (b) Histogram of mIoU_{prop} for different proposal generation methods. mIoU_{prop} denotes the mean IoU between proposal boxes and ground-truth for an object. Small mIoU_{prop} in MCG brings semantic confusion. Whereas for our P2BNet with refinement, large mIoU_{prop} is beneficial for optimization. Statistics are on COCO-17 training set, and both figures have 50 bins.

too many proposal boxes that do not have any intersection with any of the foreground objects; 2) Positive proposals per object are unbalanced. The positive proposals per object produced by MCG on the COCO-17 training set are shown in Fig. 2(a), which is clearly off-balance; 3) Majority of the proposals in bags have very low IoU indicating low-quality proposals (Fig. 2(b)). Also, as the previous PSOD methods only construct image-level bags, they can not utilize the point annotations during MIL training leading to a mixture of different objects in the same bag. All these problems limit the overall quality of the constructed bags, which contributes to the poor performance of the model.

In this paper, we propose P2BNet as an alternative to the OTSP methods for generating high-quality object proposals. The number of proposals generated by P2BNet is balanced for each object, and they cover varied scales and aspect ratios. Additionally, the proposal bags are instance-level instead of image-level. This preserves the exclusivity of objects for a given proposal bag which is very helpful during MIL training. To further improve the quality of the bag, a coarse-to-fine procedure is designed in a cascade fashion in P2BNet. The refinement stage consists of two parts, the coarse pseudo-box prediction (CBP) and the precise pseudo-box refinement (PBR). The CBP stage predicts the coarse scale (width and height) of objects, whereas the PBR stage iteratively finetunes the scale and position. Our P2BNet generates high-quality, balanced proposal bags and ensures the contribution of point annotations in all stages (before, during, and after MIL training). The detailed experiments on COCO suggest the effectiveness and robustness of our model outperforming the previous point-based detectors by a large margin. Our main contributions are as follows:

— P2BNet, a generative and OTSP-free network, is designed for predicting pseudo boxes. It generates inter-objects balanced instance-level bags and is beneficial for better optimization of MIL training. In addition, P2BNet is much more time-efficient than the OTSP-base methods.

— A coarse-to-fine fashion in P2BNet with CBP and PBR stage is proposed for higher-quality proposal bags and better prediction.
— The detection performance of our proposed P2BNet-FR framework with P2BNet under single quasi-center point supervision improves the mean average precision (AP) of the previous best PSOD method by more than 50% (relative) on COCO and bridges the gap between bounding box supervised detectors achieving comparable performance on AP_{50}.

2 Related Work

In this section, we briefly discuss the research status of box-supervised, image-level and point-level supervised object detection.

2.1 Box-Supervised Object Detection

Box-supervised object detection [13,30,29,25,23,4,38,46] is a traditional object detection paradigm that gives the network a specific category and box information. One-stage detectors based on sliding-window, like YOLO [29], SSD [25], and RetinaNet [23], predict classification and bounding-box regression through setting anchors. Two-stage detectors predict proposal boxes through OTSP methods (like selective search [34] in Fast R-CNN [13]) or deep networks (like RPN in Faster R-CNN [30]) and conduct classification and bounding-box regression with filtered proposal boxes sparsely. Transformer-based detectors (DETR [4], Deformable-DETR [52], and Swin-Transformer [26]) come, utilizing global information for better representation. Sparse R-CNN [38] combines the advantages of transformer and CNN to a sparse detector. [43,9,14] study on oriented object detection in aerial scenario. However, box-level annotation requires high costs.

2.2 Image-Supervised Object Detection

Image-supervised object detection [2,39,40,6,41,51,8,49,48,27,35] is the traditional field in WSOD. The traditional image-supervised WSOD methods can be divided into two styles: MIL-based [2,39,40,6,41], and CAM-based [51,8,49].

In MIL-based methods, a bag is positively labelled if it contains at least one positive instance; otherwise, it is negative. The objective of MIL is to select positive instances from a positive bag. WSDDN [2] introduced MIL into WSOD with a representative two-stream weakly supervised deep detection network that can classify positive proposals. OICR [39] introduces iterative fashion into WSOD and attempts to find the whole part instead of a discriminative part. PCL [40] develops the proposal cluster learning and uses the proposal clusters as supervision to indicate the rough locations where objects most likely appear. Subsequently, SLV [6] brings in spatial likelihood voting to replace the max score proposal, further looking for the whole context of objects. Our paper produces the anchor-like [35,30] proposals around the point annotation as a bag and uses

instance-level MIL to train the classifier. It moves the fixed pre-generated proposals (*e.g.*OICR, PCL and UWSOD [35]) to achieve the coarse to fine purpose.

In CAM-based methods, the main idea is to produce the class activation maps (CAM) [51], use threshold to choose a high score region, and find the smallest circumscribed rectangle of the largest general domain. WCCN [8] uses a three-stage cascade structure. The first stage produces the class activation maps and obtains the initial proposals, the second stage is a segmentation network for refining object localization, and the last stage is a MIL stage outputting the results. Acol [49] introduces two parallel-classifiers for object localization using adversarial complementary learning to alleviate the discriminative region.

2.3 Point-Supervised Object Detection

Point-level annotation is a fairly recent innovation. The average time for annotating a single point is about 1.87s per image, close to image-level annotation (1.5 s/image) and much lower than that for bounding box(34.5 s/image). The statistics [11,28] are performed on VOC [10], which can be analogized to COCO [24].

[28] introduces center-click annotation to replace box supervision and estimates scale with the error between two times of center-click. [32] designs a network compatible with various supervision forms like tags, points, scribbles, and boxes annotation. However, these frameworks are based on OTSP methods and are not specially designed for point annotation. Therefore, the performance is limited and performs poorly in complex scenarios like the COCO [24] dataset. We introduce a new framework with P2BNet which is free of OTSP methods.

3 Point-to-Box Network

The P2BNet-FR framework consists of Point-to-Box Network (P2BNet) and Faster R-CNN (FR). P2BNet predicts pseudo boxes with point annotations to train the detector. We use standard settings for Faster R-CNN without any bells and whistles. Hence, we go over the proposed P2BNet in detail in this section.

The architecture of P2BNet is shown in Fig. 3, which includes the coarse pseudo box prediction (CBP) stage and the pseudo box refinement (PBR) stage. The CBP stage predicts the coarse scale (width and height) of objects, whereas the PBR stage iteratively finetunes the scale and position. The overall loss function of P2BNet is the summation of the losses of these two stages, *i.e.*,

$$\mathcal{L}_{p2b} = \mathcal{L}_{cbp} + \sum_{t=1}^{T} \mathcal{L}_{pbr}^{(t)}, \tag{1}$$

where PBR includes T iterations, and $\mathcal{L}_{pbr}^{(t)}$ is the loss of t-th iteration.

3.1 Coarse Pseudo Box Prediction

In the CBP stage, firstly, proposal boxes of different widths and heights are generated in an anchor-style for each object, taking the annotated point as the

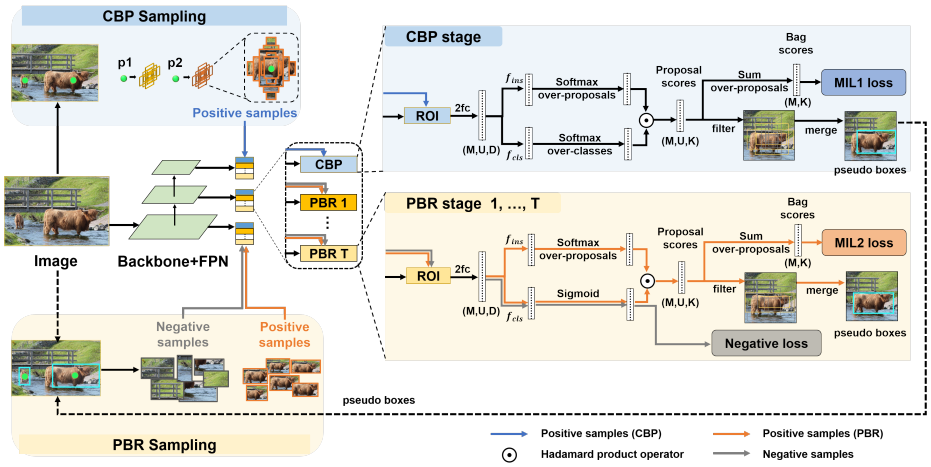


Fig. 3. The architecture of P2BNet. Firstly, to predict coarse pseudo boxes in CBP stage, proposal bags are fixedly sampled around point annotations for classifier training. Then, to predict refined pseudo boxes in PBR stage, high-quality proposal bags and negative proposals are sampled with coarse pseudo boxes for training. Finally, the pseudo boxes generated by the trained P2BNet serve as supervision for the training the classic detector. (Best viewed in color.)

box center. Secondly, features of the sampled proposals are extracted to train a MIL classifier for selecting the best fitted proposal of objects. Finally, the top-k merging policy are utilized to estimate coarse pseudo boxes.

CBP Sampling: fixed sampling around the annotated point. With the point annotation $p = (p_x, p_y)$ as the center, s as the size, and v to adjust the aspect ratio, the proposal box $b = (b_x, b_y, b_w, b_h)$ is generated, *i.e.* $b = (p_x, p_y, v \cdot s, \frac{1}{v} \cdot s)$. The schematic diagram of proposal box sampling is shown in Fig. 4 (Left). By adjusting s and v, each point annotation p_j generates a bag of proposal boxes with different scales and aspect ratios, denoted by $\mathcal{B}_j$ ($j \in \{1, 2, \ldots, M\}$, where M is the amount of objects). The details of the settings of s and v are given in supplemental. All proposal bags are utilized for training the MIL classifier in the CBP module with the category labels of points as supervision.

There is a minor issue that oversized s may lead most of b outside the image and introduce too many meaningless padding values. In this case, we clip b to guarantee that it is inside the image (see Fig. 4 (Left)), *i.e.*,

$$b = \left(p_x, p_y, min(v \cdot s, 2(p_x - 0), 2(W - p_x)), min(\frac{1}{v} \cdot s, 2(p_y - 0), 2(H - p_y)) \right), \quad (2)$$

where W and H denote the image size. $(p_x - 0)$ and $(W - p_x)$ are the distances from the center to the left and right edges of the image, respectively.

CBP Module. For a proposal bag $\mathcal{B}_j$, features $\mathbf{F}_j \in \mathbb{R}^{U \times D}$ are extracted through 7×7 RoIAlign [15] and two fully connected (fc) layers, where U is the number of proposals in $\mathcal{B}_j$, and D is the feature dimension. We refer to WS-

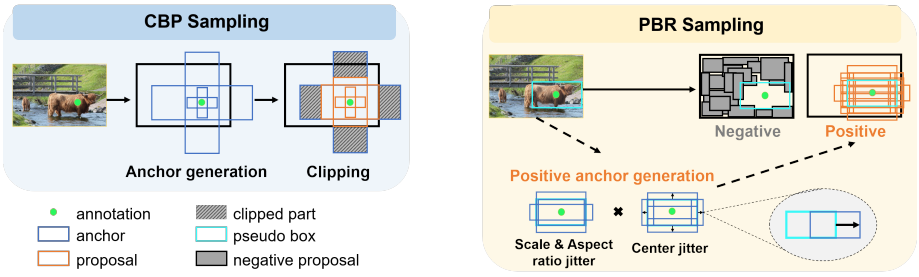


Fig. 4. Details of sampling strategies in the CBP stage and the PBR stage. The arrows in PBR sampling mean the offset of center jitter. Samples are obtained through center jitter following scale and aspect ratio jatter in PBR sampling.

DDN [2] and design a two-stream structure as a MIL classifier to find the best bounding box region to represent the object. Specifically, applying the classification branch f_{cls} to $\mathbf{F}_j$ yields $\mathbf{O}_j^{cls} \in \mathbb{R}^{U\times K}$, which is then passed through the activation function to obtain the classification score $\mathbf{S}_j^{cls} \in \mathbb{R}^{U\times K}$, where K represents the number of instance categories. Likewise, instance score $\mathbf{S}_j^{ins} \in \mathbb{R}^{U\times K}$ is obtained through instance selection branch f_{ins} and activation function, *i.e.*,

$$\mathbf{O}_j^{cls} = f_{cls}(\mathbf{F}_j), \quad [\mathbf{S}_j^{cls}]_{uk} = e^{[\mathbf{O}_j^{cls}]_{uk}} \Big/ \sum_{i=1}^{K} e^{[\mathbf{O}_j^{cls}]_{ui}}; \tag{3}$$

$$\mathbf{O}_j^{ins} = f_{ins}(\mathbf{F}_j), \quad [\mathbf{S}_j^{ins}]_{uk} = e^{[\mathbf{O}_j^{ins}]_{uk}} \Big/ \sum_{i=1}^{U} e^{[\mathbf{O}_j^{ins}]_{ik}}, \tag{4}$$

where $[\cdot]_{uk}$ denotes the value at row u and column k in the matrix. The proposal score $\mathbf{S}_j$ is obtained by computing the Hadamard product of the classification score and the instance score, and the bag score $\widehat{\mathbf{S}}_j$ is obtained by the summation of the proposal scores of U proposal boxes, *i.e.*,

$$\mathbf{S}_j = \mathbf{S}_j^{cls} \odot \mathbf{S}_j^{ins} \in \mathbb{R}^{U\times K}, \quad \widehat{\mathbf{S}}_j = \sum_{u=1}^{U} [\mathbf{S}_j]_u \in \mathbb{R}^{K}. \tag{5}$$

$\widehat{\mathbf{S}}_j$ can be seen as the weighted summation of the classification score $[\mathbf{S}_j^{cls}]_u$ by the corresponding selection score $[\mathbf{S}_j^{ins}]_u$.

CBP Loss. The MIL loss in the CBP module (termed $\mathcal{L}_{mil1}$ to distinguish it from the MIL loss in PBR) uses the form of cross-entropy loss, defined as:

$$\mathcal{L}_{cbp} = \alpha_{mil1}\mathcal{L}_{mil1} = -\frac{\alpha_{mil1}}{M}\sum_{j=1}^{M}\sum_{k=1}^{K}[\mathbf{c}_j]_k \log([\widehat{\mathbf{S}}_j]_k) + (1-[\mathbf{c}_j]_k)\log(1-[\widehat{\mathbf{S}}_j]_k), \tag{6}$$

where $\mathbf{c}_j \in \{0,1\}^K$ is the one-hot category label, α_{mil1} is 0.25. The CBP loss is to make each proposal correctly predict the category and instance it belongs to.

Finally, the top-k boxes with the highest proposal scores $\mathbf{S}_j$ of each object are weighted to obtain coarse pseudo boxes for the following PBR sampling.

3.2 Pseudo Box Refinement

The PBR stage aims to finetune the position, width and height of pseudo boxes, and it can be performed iteratively in a cascaded fashion for better performance. By adjusting the height and width of the pseudo box obtained in the previous stage (or iteration) in a small span while jittering its center position, finer proposal boxes are generated as positive examples for module training. Further, because the positive proposal bags are generated in the local region, negative samples can be sampled far from the proposal bags to suppress the background. The PBR module also weights the top-k proposals with the highest predicted scores to obtain the refined pseudo boxes, which are the final output of P2BNet.

PBR Sampling. adaptive sampling around estimated boxes. As shown in Fig. 4 (Right), for each coarse pseudo box $b^* = (b_x^*, b_y^*, b_w^*, b_h^*)$ obtained in the previous stage (or iteration), we adjust its scale and aspect ratio with s and v, and jitter its postion with o_x, o_y to obtain the finer proposal $b = (b_x, b_y, b_w, b_h)$:

$$b_w = v \cdot s \cdot b_w^*, \quad b_h = \frac{1}{v} \cdot s \cdot b_h^*, \tag{7}$$

$$b_x = b_x^* + b_w \cdot o_x, \quad b_y = b_y^* + b_h \cdot o_y. \tag{8}$$

These finer proposals are used as positive proposal bag $\mathcal{B}_j$ to train PBR module.

Furthermore, to better suppress the background, negative samples are introduced in the PBR sampling. We randomly sample many proposal boxes, which have small IoU (by default set as smaller than 0.3) with all positive proposals in all bags, to compose the negative sample set $\mathcal{N}$ for the PBR module. Through sampling proposal boxes by pseudo box distribution, high-quality proposal boxes are obtained for better optimization (shown in Fig. 5).

PBR Module. The PBR module has a similar structure to the CBP module. It shares the backbone network and two fully connected layers with CBP, and also has a classification branch f_{cls} and an instance selection branch f_{ins}. Note that f_{cls} and f_{ins} do not share parameters between different stages and iterations. For instance selection branch, we adopt the same structure as the CBP module, and utilize Eq. 4 to predict the instance score $\mathbf{S}_j^{ins}$ for the proposal bag $\mathcal{B}_j$. Differently, the classification branch uses the *sigmoid* activation function $\sigma(x)$ to predict the classification score $\mathbf{S}_j^{cls}$, *i.e.*,

$$\sigma(x) = 1/(1+e^{-x}), \quad \mathbf{S}_j^{cls} = \sigma(f_{cls}(\mathbf{F}_j)) \in \mathbb{R}^{U \times K}. \tag{9}$$

This form makes it possible to perform multi-label classification, which can distinguish overlapping proposal boxes from different objects. According to the form of Eq. 5, bag score $\widehat{\mathbf{S}}_j^*$ is calculated using $\mathbf{S}_j^{cls}$ and $\mathbf{S}_j^{ins}$ of the current stage.

For the negative sample set $\mathcal{N}$, we calculate its classification score as:

$$\mathbf{S}_{neg}^{cls} = \sigma(f_{cls}(\mathbf{F}_{neg})) \in \mathbb{R}^{|\mathcal{N}| \times K}. \tag{10}$$

PBR Loss. The PBR loss consists of MIL loss $\mathcal{L}_{mil2}$ for positive bags and negative loss $\mathcal{L}_{neg}$ for negative samples, *i.e.*,

$$\mathcal{L}_{pbr} = \alpha_{mil2}\mathcal{L}_{mil2} + \alpha_{neg}\mathcal{L}_{neg}, \tag{11}$$

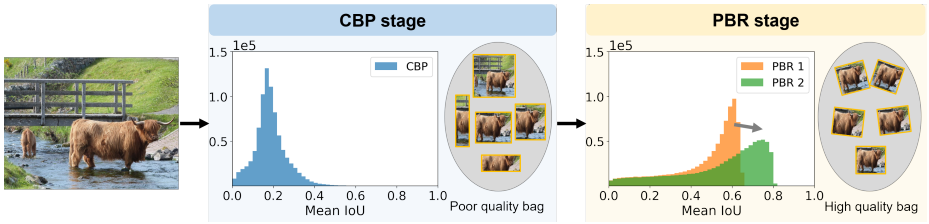


Fig. 5. The progression of the mIoU_{prop} during refinement. By statistics, the mIoU_{pred} is gradually increasing in the PBR stage, indicating that the quality of the proposal bag improves in iterative refinement.

where $\alpha_{mil2} = 0.25$ and $\alpha_{neg} = 0.75$ are the settings in this paper.

1) MIL Loss. The MIL loss $\mathcal{L}_{mil2}$ in the PBR stage is defined as:

$$\text{FL}(\zeta, \tau) = -\sum_{k=1}^{K} [\tau]_k (1 - [\zeta]_k)^{\gamma} \log([\zeta]_k) + (1 - [\tau]_k)([\zeta]_k)^{\gamma} \log(1 - [\zeta]_k), \tag{12}$$

$$\mathcal{L}_{mil2} = \frac{1}{M} \sum_{j=1}^{M} \left\langle \mathbf{c}_j^{\mathrm{T}}, \widehat{\mathbf{S}}_j^{*} \right\rangle \cdot \text{FL}(\widehat{\mathbf{S}}_j, \mathbf{c}_j), \tag{13}$$

where $\text{FL}(\zeta, \tau)$ is the focal loss [23], and γ is set as 2 following [23]. $\widehat{\mathbf{S}}_j^{*}$ represents the bag score of the last PBR iteration (for the first iteration of PBR, using the bag score in CBP). $\left\langle \mathbf{c}_j^{\mathrm{T}}, \widehat{\mathbf{S}}_j^{*} \right\rangle$ represents the inner product of the two vectors, which means the predicted bag score of the previous stage or iteration on ground-truth category. Score is used to weight the FL of each object for stable training.

2) Negative Loss. Conventional MIL treats proposal boxes belonging to other categories as negative samples. In order to further suppress the backgrounds, we sample more negative samples in the PBR stage and introduce the negative loss (γ is also set to 2 following FL), *i.e.*,

$$\beta = \frac{1}{M} \sum_{j=1}^{M} \left\langle \mathbf{c}_j^{\mathrm{T}}, \widehat{\mathbf{S}}_j^{*} \right\rangle, \quad \mathcal{L}_{neg} = -\frac{1}{|\mathcal{N}|} \sum_{\mathcal{N}} \sum_{k=1}^{K} \beta \cdot ([\mathbf{S}_{neg}^{cls}]_k)^{\gamma} \log(1 - [\mathbf{S}_{neg}^{cls}]_k). \tag{14}$$

4 Experiments

4.1 Experiment Settings

Datasets and Evaluate Metrics. For experiments, we use the public available MS COCO [24] dataset. COCO has 80 different categories and two versions. COCO-14 has 80K training and 40K validation images whereas COCO-17 has

118K training and 5K validation images. Since the ground truth on the test set is not released, we train our model on the training set and evaluate it on the validation set reporting AP_{50} and AP (averaged over IoU thresholds in $[0.5 : 0.05 : 0.95]$) on COCO. The $mIoU_{pred}$ is calculated by the mean IoU between predicted pseudo boxes and their corresponding ground-truth bounding-boxes of all objects in the training set. It can directly evaluate the ability of P2BNet to transform annotated points into accurate pseudo boxes.

Implementation Details. Our codes of P2BNet-FR are based on MMDetection [5]. The stochastic gradient descent (SGD [3]) algorithm is used to optimize in $1\times$ training schedule. The learning rate is set to 0.02 and decays by 0.1 at the 8-th and 11-th epochs, respectively. In P2BNet, we use multi-scale (480, 576, 688, 864, 1000, 1200) as the short side to resize the image during training and single-scale (1200) during inference. We choose the classic Faster R-CNN FPN [30,22] (backbone is ResNet-50 [16]) as the detector with the default setting, and single-scale (800) images are used during training and inference. More details are included in the supplementary section.

Quasi-Center Point Annotation. We propose a quasi-center (QC) point annotation that is friendly for object detection tasks with a low cost. In practical scenarios, we ask annotators to annotate the object in the non-high limit center region with a loose rule. Since datasets in the experiment are already annotated with bounding boxes or masks, it is reasonable that the manually annotated points follow Gaussian distribution in the central region. We utilize Rectified Gaussian Distribution (RG) defined in [45] with central ellipse constraints. For a bounding box of $b = (b_x, b_y, b_w, b_h)$, its central ellipse can be defined as $Ellipse(\kappa)$, using (b_x, b_y) as the ellipse center and $(\kappa \cdot b_w, \kappa \cdot b_h)$ as the two axes of the ellipse. In addition, in view of the fact that the absolute position offset for a large object is too large under the above rule, we limit the two axes to no longer than 96 pixels. If the object's mask $Mask$ overlaps with the central ellipse $Ellipse(\kappa)$, V is used to denote the intersection. If there is no intersecting area, V represents the entire $Mask$. When generated from bounding box annotations, the boxes are treated as masks. Then RG is defined as,

$$RG(p; \mu, \sigma, \kappa) = \begin{cases} \frac{Gauss(p;\mu,\sigma)}{\int_V Gauss(p;\mu,\sigma)dp}, & p \in V \\ 0, & p \notin V \end{cases} \tag{15}$$

where μ and σ are mean and standard deviation of RG. κ decides the $Ellipse(\kappa)$. In this paper, $RG(p; 0, \frac{1}{4}, \frac{1}{4})$ is chosen to generate the QC point annotations.

4.2 Performance Comparisons

Unless otherwise specified, the default components of our P2BNet-FR framework are P2BNet and Faster R-CNN. We compare the P2BNet-FR with the existing PSOD methods while choosing the state-of-the-art UFO^2 [32] framework as the baseline for comprehensive comparisons. In addition, to demonstrate the performance advantages of the PSOD methods, we compare them with the state-of-the-art WSOD methods. At the same time, we compare the performance of the box-supervised object detectors to reflect their performance upper bound.

Method	Backbone	Proposal	COCO-14		COCO-17	
			AP	AP_{50}	AP	AP_{50}
Box-supervised detectors						
Fast R-CNN [13]	VGG-16	SS	18.9	38.6	19.3	39.3
Faster R-CNN [30]	VGG-16	RPN	21.2	41.5	21.5	42.1
FPN [5]	R-50	RPN	**35.5**	**56.7**	37.4	**58.1**
RetinaNet [23,5]	R-50	-	34.3	53.3	36.5	55.4
Reppoint [44,5]	R-50	-	-	-	37.0	56.7
Sparse R-CNN [38,5]	R-50	PP	-	-	**37.9**	56.0
Image-supervised detectors						
OICR+Fast [39,13]	VGG-16	SS	7.7	17.4	-	-
PCL [40]	VGG-16	SS	8.5	19.4	-	-
PCL+Fast [40,13]	VGG-16	SS	9.2	19.6	-	-
MEFF+Fast [12,13]	VGG-16	SS	8.9	19.3	-	-
C-MIDN [42]	VGG-16	SS	9.6	21.4	-	-
WSOD2 [47]	VGG-16	SS	10.8	22.7	-	-
UFO^{2*} [32]	VGG-16	MCG	10.8	23.1	-	-
GradingNet-C-MIL [18]	VGG-16	SS	11.6	25.0	-	-
ICMWSD [31]	VGG-16	MCG	11.4	24.3	-	-
ICMWSD [31]	R-50	MCG	12.6	26.1	-	-
ICMWSD [31]	R-101	MCG	13.0	26.3	-	-
CASD [17]	VGG-16	SS	12.8	26.4	-	-
CASD [17]	R-50	SS	**13.9**	**27.8**	-	-
Point-supervised detectors						
Click [28]	AlexNet	SS	-	18.4	-	-
UFO^2 [32]	VGG-16	MCG	12.4	27.0	-	-
$UFO^{2\dagger}$ [32]	VGG-16	MCG	12.8	26.6	13.2	27.2
$UFO^{2\ddagger}$ [32]	VGG-16	MCG	12.7	26.5	13.5	27.9
$UFO^{2\ddagger}$ [32]	R-50	MCG	12.6	27.6	13.2	28.9
P2BNet-FR (Ours)	R-50	Free	**19.4**	**43.5**	**22.1**	**47.3**

Table 1. The performance comparison of box-supervised, image-supervised, and point-supervised detectors on COCO dataset. * means UFO^2 with image-level annotation. † means the performance we reproduce with the original setting. ‡ means we re-implement UFO^2 with our QC point annotation. The performance of P2BNet-FR, UFO^2, and the box-supervised detector is tested on a single scale dataset. Our P2BNet-FR is based on P2BNet with top-4 merging and one PBR stage. SS is selective search [34], PP means proposal box defined in [38], and Free represents OTSP-free based method.

Comparison with PSOD Methods. We compare the existing PSOD methods Click [28] and UFO^2 [32] on COCO, as shown in Tab. 1. Both Click and UFO^2 utilize OTSP-based methods (SS [34] or MCG [1]) to generate proposal boxes. Since the point annotation used by UFO^2 is different from the QC point proposed in this paper, for a fair comparison, we re-train UFO^2 on the public code with our QC point annotation. In addition, the previous methods are mainly based on VGG-16 [36] or AlexNet [20]. For consistency, we extend the UFO^2 to the ResNet-50 FPN backbone and compare it with our framework. In comparison with Click and UFO^2, our P2BNet-FR framework outperforms them by a large margin. On COCO-14, P2BNet-FR improves AP and AP_{50} by 6.8 and 15.9, respectively. Also, our framework significantly outperforms state-of-the-art performance by 8.9 AP and 18.4 AP_{50} on COCO-17. In Fig. 6, the visualization

CBP stage		PBR stage			Performance		
$\mathcal{L}_{pos}$	$\mathcal{L}_{mil1}$	$\mathcal{L}_{mil2}$	$\mathcal{L}_{neg}$	$\mathcal{L}_{pesudo}$	mIoU_{pred}	AP	AP_{50}
✓					25.0	2.9	10.3
	✓				50.2	13.7	37.8
	✓	✓			52.0	12.7	35.4
	✓	✓	✓		**57.4**	**21.7**	**46.1**
	✓	✓	✓	✓	56.7	18.5	44.1

(a) The effectiveness of training loss in P2BNet: $\mathcal{L}_{mil1}$ in CBP stage, $\mathcal{L}_{mil2}$ and $\mathcal{L}_{neg}$ in PBR stage. $\mathcal{L}_{pos}$ and $\mathcal{L}_{pesudo}$ is for comparison.

top-k	mIoU_{pred}	AP	AP_{50}
1	49.2	12.2	35.9
3	54.7	21.3	46.6
4	**57.5**	**22.1**	**47.3**
7	57.4	21.7	46.1
10	57.1	21.5	46.0

(b) The top-k policy for box merging. k is set the same for all stages.

T	mIoU_{pred}	AP	AP_{50}
0	50.2	13.7	37.8
1	**57.4**	21.7	**46.1**
2	57.0	**21.9**	**46.1**
3	56.2	21.3	45.6

(c) The number of iterations T in the PBR stage. $T = 0$ means only the CBP stage is conducted.

Table 2. Ablation study (Part I).

shows our P2BNet-FR makes full use of the precise location information of point annotation and can distinguish dense objects in complex scenes.

Comparison with WSOD Methods. We compare the proposed framework to the state-of-the-art WSOD methods on the COCO-14 in Tab. 1. The performance of P2BNet-FR proves that compared with WSOD, PSOD significantly improves the detection performance with little increase in the annotation cost, showing that the PSOD task has great prospects for development.

Comparison with Box-Supervised Methods. In order to verify the feasibility of P2BNet-FR in practical applications and show the upper bound under this supervised manner, we compare the box-supervised detector [30] in Tab. 1. Under AP_{50}, P2BNet-FR-R50 (47.3 AP_{50}) is much closer to box-supervised detector FPN-R50 (58.1 AP_{50}) than previous WSOD and PSOD method. It shows that PSOD can be applied in industries that are less demanding on box quality and more inclined to find objects [19,50], with greatly reduced annotation cost.

4.3 Ablation Study

In this section, all the ablation studies are conducted on the COCO-17 dataset. The top-k setting is $k = 7$ except for the box merging policy part in Tab. 2(b) and different detectors part ($k = 4$) in Tab. 3(d).

Training Loss in P2BNet. The ablation study of the training loss in P2BNet is shown in Tab. 2(a). **1) CBP loss.** Only with $\mathcal{L}_{mil1}$ in the CBP stage, we can obtain 13.7 AP and 37.8 AP_{50}. For comparison, we conduct $\mathcal{L}_{pos}$, which views all the proposal boxes in the bag as positive samples. We find it hard to optimize, and the performance is bad, demonstrating the effectiveness of our proposed $\mathcal{L}_{mil1}$ for pseudo box prediction. Coarse proposal bags can cover most objects in high IoU, resulting in a low missing rate. However, the performance still has

Methods	AR_1	AR_{10}	AR_{100}
UFO^2	14.7	22.6	23.3
P2BNet-FR	**21.3**	**32.8**	**34.2**

(a) Comparisons of average recall for UFO^2 and P2BNet-FR.

Balance	AP	AP_{50}
✓	**21.7**	**46.1**
-	12.9	36.0

(b) Unbalance issue.

Jitter	AP	AP_{50}
✓	**21.7**	**46.1**
-	14.2	38.2

(c) Jitter strategy.

Detectors	GT box		Pseudo box	
	AP	AP_{50}	AP	AP_{50}
RetinaNet [23]	36.5	55.4	21.0	44.9
Reppoint [44]	37.0	56.7	20.8	45.1
Sparse R-CNN [38]	**37.9**	56.0	21.1	43.3
FR-FPN [30,22]	37.4	**58.1**	**22.1**	**47.3**

(d) Performance of different detectors on ground-truth box annotations and pseudo boxes generated by P2BNet. We use the top-4 for box merging.

Table 3. Ablation study (Part II).

the potential to be refined because the scale and aspect ratio are coarse, and the center position needs adjustment. **2) PBR loss.** With a refined sampling of proposal bag (shown in Fig. 5), corresponding PBR loss is introduced. Only with $\mathcal{L}_{mil2}$, the performance is just 12.7 AP. The main reasons of performance degradation are error accumulation in a cascade fashion and lacking negative samples for focal loss. There are no explicit negative samples to suppress background for *Sigmoid* activation function, negative sampling and negative loss $\mathcal{L}_{neg}$ is introduced. Performance increases by 9.0 AP and 10.7 AP_{50}, indicating that it is essential and effectively improves the optimization. We also evaluate the $mIoU_{pred}$ to discuss the predicted pseudo box's quality. In the PBR stage with $\mathcal{L}_{mil2}$ and $\mathcal{L}_{neg}$, the mIoU increases from 50.2 to 57.4, suggesting better quality of the pseudo box. Motivated by [45], we conduct $\mathcal{L}_{pesudo}$, viewing pseudo boxes from the CBP stage as positive samples. However, the $\mathcal{L}_{pesudo}$ limits the refinement and the performance decreases. In Tab. 3(c), if we remove the jitter strategy of proposal boxes in PBR stage, the performance drops to 14.2 AP.

Number of Refinements in PBR. Refining pseudo boxes is a vital part of P2BNet, and the cascade structure is used for iterative refinement to improve performance. Tab. 2(c) shows the effect of the refining number in the PBR stage. One refinement brings a performance gain of 8.0 AP, up to a competitive 21.7 AP. The highest 21.9 AP is obtained with two refinements, and the performance is saturated. We choose one refinement as the default configuration.

Box Merging Policy. We use the top-k score average weight as our merging policy. We find that the hyper-parameter k is slightly sensitive and can be easily generalized to other datasets, as presented in Tab. 2(b), and only the top-1 or top-few proposal box plays a leading role in box merging. The best performance is 22.1 AP and 47.3 AP_{50} when $k = 4$. The $mIoU_{pred}$ between the pseudo box and ground-truth box is 57.5. In inference, if bag score $\mathbf{S}$ is replaced by classification score $\mathbf{S}^{cls}$ for merging, the performance drops to 17.4 AP (vs 21.7 AP).

Average Recall. In Tab. 3(a), the AR in UFO^2 is 23.3, indicating a higher missing rate. Whereas the P2BNet-FR obtains 34.2 AR, far beyond that of the UFO^2. It shows our OTSP-free method is better at finding objects.

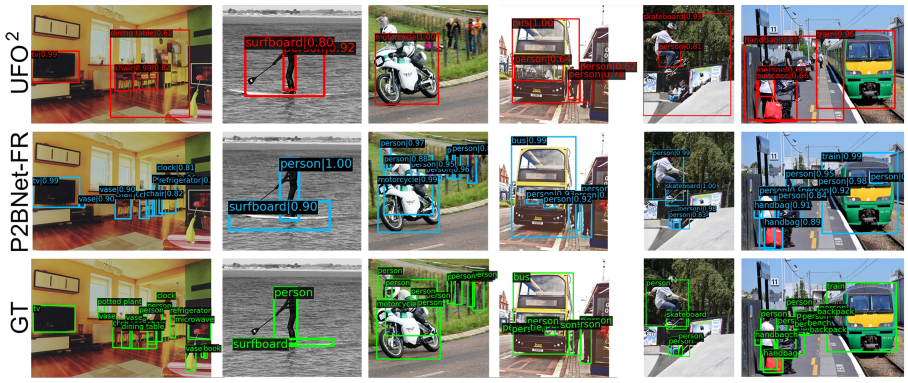


Fig. 6. Visualization of detection results of P2BNet-FR and UFO^2. Our P2BNet-FR can distinguish dense objects and perform well in complex scene. (Best viewed in color.)

Unbalance Sampling Analysis. To demonstrate the effect of unbalance sampling, we sample different numbers of proposal boxes for each object and keep them constant in every epoch during the training period. The performance drops in Tab. 3(b) suggests the negative impact of unbalanced sampling.

Different Detectors. We train different detectors [30,22,23,44,38] for the integrity experiments, all of which are conducted on R-50, as shown in Tab. 3(d). Our framework exhibits competitive performance on other detectors. Box supervised performances are listed to demonstrate the upper bound of our framework.

5 Conclusion

In this paper, we give an in-depth analysis of shortcomings in OTSP-based PSOD frameworks, and further propose a novel OTSP-free network termed P2BNet to obtain inter-objects balanced and high-quality proposal bags. The coarse-to-fine strategy divides the prediction of pseudo boxes into CBP and PBR stages. In the CBP stage, fixed sampling is performed around the annotated points, and coarse pseudo boxes are predicted through instance-level MIL. The PBR stage performs adaptive sampling around the estimated boxes to finetune the predicted boxes in a cascaded fashion. As mentioned above, P2BNet takes full advantage of point information to generate high-quality proposal bags, which is more conducive to optimizing the detector (FR). Remarkably, the conceptually simple P2BNet-FR framework yields state-of-the-art performance with single point annotation.

Acknowledgements This work was supported in part by the Youth Innovation Promotion Association CAS, the National Natural Science Foundation of China (NSFC) under Grant No. 61836012, 61771447 and 62006244, the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No.XDA27000000, and Young Elite Scientist Sponsorship Program of China Association for Science and Technology YESS20200140.

References

1. Arbeláez, P.A., Pont-Tuset, J., *et al.*, J.T.B.: Multiscale combinatorial grouping. In: CVPR (2014) 2, 11
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: CVPR (2016) 2, 4, 7
3. Bottou, L.: Stochastic gradient descent tricks. In: Neural Networks: Tricks of the Trade - Second Edition. Springer (2012) 10
4. Carion, N., Massa, F., *et al.*, G.S.: End-to-end object detection with transformers. In: ECCV (2020) 1, 4
5. Chen, K., Wang, J., Pang, J.e.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019) 10, 11
6. Chen, Z., Fu, Z., *et al.*, R.J.: SLV: spatial likelihood voting for weakly supervised object detection. In: CVPR (2020) 2, 4
7. Cheng, B., Parkhi, O., Kirillov, A.: Pointly-supervised instance segmentation. CoRR (2021) 2
8. Diba, A., Sharma, V., *et al.*, A.M.P.: Weakly supervised cascaded convolutional networks. In: CVPR (2017) 2, 4, 5
9. Ding, J., Xue, N., Long, Y., Xia, G., Lu, Q.: Learning roi transformer for oriented object detection in aerial images. In: CVPR (2019) 4
10. Everingham, M., Gool, L.V., *et al.*, C.K.I.W.: The pascal visual object classes (VOC) challenge. IJCV (2010) 5
11. Gao, M., Li, A., *et al.*, R.Y.: C-WSL: count-guided weakly supervised localization. In: ECCV (2018) 5
12. Ge, W., Yang, S., Yu, Y.: Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In: CVPR (2018) 11
13. Girshick, R.B.: Fast R-CNN. In: ICCV (2015) 1, 4, 11
14. Guo, Z., Liu, C., Zhang, X., Jiao, J., Ji, X., Ye, Q.: Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In: CVPR (2021) 4
15. He, K., Gkioxari, G., *et al.*, P.D.: Mask R-CNN. In: ICCV (2017) 6
16. He, K., Zhang, X., *et al.*, S.R.: Deep residual learning for image recognition. In: CVPR (2016) 10
17. Huang, Z., Zou, Y., *et al.*, B.V.K.V.K.: Comprehensive attention self-distillation for weakly-supervised object detection. In: NeurIPS (2020) 11
18. Jia, Q., Wei, S., *et al.*, T.R.: Gradingnet: Towards providing reliable supervisions for weakly supervised object detection by grading the box candidates. In: AAAI (2021) 11
19. Jiang, N., Wang, K., Peng, X., Yu, X., Wang, Q., Xing, J., Li, G., Zhao, J., Guo, G., Han, Z.: Anti-uav: A large multi-modal benchmark for UAV tracking. IEEE TMM (2021) 12
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012) 11
21. Lee, P., Byun, H.: Learning action completeness from points for weakly-supervised temporal action localization. In: ICCV (2021) 2
22. Lin, T., Dollár, P., *et al.*, R.B.G.: Feature pyramid networks for object detection. In: CVPR (2017) 10, 13, 14
23. Lin, T., Goyal, P., *et al.*, R.B.G.: Focal loss for dense object detection. In: ICCV (2017) 1, 4, 9, 11, 13, 14

24. Lin, T.Y., Maire, M.e.: Microsoft coco: Common objects in context. In: ECCV (2014) 5, 9
25. Liu, W., Anguelov, D., *et al.*, D.E.: SSD: single shot multibox detector. In: ECCV (2016) 1, 4
26. Liu, Z., Lin, Y., *et al.*, Y.C.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021) 4
27. Meng, M., Zhang, T., Yang, W., Zhao, J., Zhang, Y., Wu, F.: Diverse complementary part mining for weakly supervised object localization. IEEE TIP (2022) 4
28. Papadopoulos, D.P., Uijlings, J.R.R., *et al.*, F.K.: Training object class detectors with click supervision. In: CVPR (2017) 2, 5, 11
29. Redmon, J., Divvala, S.K., *et al.*, R.B.G.: You only look once: Unified, real-time object detection. In: CVPR (2016) 1, 4
30. Ren, S., He, K., *et al.*, R.B.G.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE TPAMI (2017) 1, 4, 10, 11, 12, 13, 14
31. Ren, Z., Yu, Z., *et al.*, X.Y.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In: CVPR (2020) 11
32. Ren, Z., Yu, Z., *et al.*, X.Y.: Ufo^2: A unified framework towards omni-supervised object detection. In: ECCV (2020) 2, 5, 10, 11
33. Ribera, J., Guera, D., Chen, Y., Delp, E.J.: Locating objects without bounding boxes. In: CVPR (2019) 2
34. van de Sande, K.E.A., Uijlings, J.R.R., *et al.*, T.G.: Segmentation as selective search for object recognition. In: ICCV (2011) 2, 4, 11
35. Shen, Y., Ji, R., Chen, Z., Wu, Y., Huang, F.: UWSOD: toward fully-supervised-level capacity weakly supervised object detection. In: NeurIPS (2020) 4, 5
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) 11
37. Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Wu, Y.: Rethinking counting and localization in crowds: A purely point-based framework. In: ICCV (2021) 2
38. Sun, P., Zhang, R., *et al.*, Y.J.: Sparse R-CNN: end-to-end object detection with learnable proposals. In: CVPR (2021) 1, 4, 11, 13, 14
39. Tang, P., *et al.*, X.W.: Multiple instance detection network with online instance classifier refinement. In: CVPR (2017) 2, 4, 11
40. Tang, P., Wang, X., *et al.*, S.B.: PCL: proposal cluster learning for weakly supervised object detection. IEEE TPAMI (2020) 2, 4, 11
41. Wan, F., Wei, P., *et al.*, Z.H.: Min-entropy latent model for weakly supervised object detection. IEEE TPAMI (2019) 2, 4
42. Yan, G., Liu, B., *et al.*, N.G.: C-MIDN: coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In: ICCV (2019) 11
43. Yang, X., Yan, J., Feng, Z., He, T.: R3det: Refined single-stage detector with feature refinement for rotating object. In: AAAI (2021) 4
44. Yang, Z., Liu, S., *et al.*, H.H.: Reppoints: Point set representation for object detection. In: ICCV (2019) 11, 13, 14
45. Yu, X., Chen, P., *et al.*, D.W.: Object localization under single coarse point supervision. In: CVPR (2022) 2, 10, 13
46. Yu, X., Gong, Y., *et al.*, N.J.: Scale match for tiny person detection. In: IEEE WACV (2020) 1, 4

47. Zeng, Z., Liu, B., *et al.*, J.F.: WSOD2: learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: ICCV (2019) 11
48. Zhang, D., Han, J., Cheng, G., Yang, M.: Weakly supervised object localization and detection: A survey. IEEE TPAMI (2021) 4
49. Zhang, X., Wei, Y., *et al.*, J.F.: Adversarial complementary learning for weakly supervised object localization. In: CVPR (2018) 2, 4, 5
50. Zhao, J., Wang, G., Li, J., Jin, L., Fan, N., Wang, M., Wang, X., Yong, T., Deng, Y., Guo, Y., Ge, S., Guo, G.: The 2nd anti-uav workshop & challenge: Methods and results. ICCVW 2021 (2021) 12
51. Zhou, B., Khosla, A., *et al.*, À.L.: Learning deep features for discriminative localization. In: CVPR (2016) 2, 4, 5
52. Zhu, X., Su, W., *et al.*, L.L.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021) 4
53. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: ECCV (2014) 2