Selecting High-Quality Proposals for Weakly Supervised Object Detection With Bottom-Up Aggregated Attention and Phase-Aware Loss

Zhihao Wu, Chengliang Liu, *Graduate Student Member, IEEE*, Jie Wen, *Member, IEEE*, Yong Xu, *Senior Member, IEEE*, Jian Yang, *Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*— Weakly supervised object detection (WSOD) has received widespread attention since it requires only image-category annotations for detector training. Many advanced approaches solve this problem by a two-phase learning framework, that is, instance mining that classifies generated proposals via multiple instance learning, and instance refinement that iteratively refines bounding boxes using the supervision produced by the preceding stage. In this paper, we observe that the detection performance is usually limited by imprecise supervision, including part domination and untight boxes. To mitigate their adverse effects, we focus on selecting high-quality proposals as the supervision for WSOD. To be specific, for the issue of part domination, we propose bottom-up aggregated attention which incorporates low-level features from shallow layers to improve location representation of top-level features. In this manner, the proposals corresponding to entire objects can get high scores. Its advantage is that it can be flexibly plugged into the WSOD framework since there is no need to attach learnable parameters or learning branches. As regards the problem of untight boxes, we propose a phase-aware loss, which is the first work to measure supervision quality by the loss in the instance mining phase, to highlight correct boxes and suppress untight ones. In this work, we unify the proposed two modules into the framework of online instance classifier refinement. Extensive experiments on the PASCAL VOC and the MS COCO demonstrate that our method can significantly improve the performance of WSOD and achieve the state-of-the-art results. The code is available at https://github.com/Horatio9702/BUAA_PALoss.**

***Index Terms*— Weakly supervised object detection, multiple instance learning, high-quality supervision.**

I. Introduction

THANKS to advanced convolutional neural networks (CNNs) [1], [2], [3], fully supervised object detection (FSOD) [4], [5], [6], [7], [8], [9], [10], [11], [12] has achieved remarkable performance. However, its training requires the large-scale dataset with bounding-box annotations [13], [14], [15], which are labor-intensive and time-consuming to be collected. To break this limitation, the community has started to train object detectors using only image-category labels, *i.e.*, weakly supervised object detection (WSOD).

The prevailing paradigm for WSOD usually adopts a two-phase learning procedure, including instance mining and instance refinement. In the first phase, the backbone network is employed to extract the features of generated proposals, and these features are exploited to classify proposals under the constraints of multiple instance learning (MIL), that is, a positive bag (*i.e.*, image) contains at least one positive instance (*i.e.*, proposal) while a negative bag only has negatives. In the second phase, multiple parallel instance classifiers are trained to refine bounding boxes using the supervision provided by the high-scoring proposals and their adjacent ones in the preceding stage.

Although this paradigm has achieved promising results, its performance is still far from that of FSOD due to inaccurate supervision. This is because there is positive feedback between supervision generation and object detection. To be specific, an accurate detection model cannot be learned without accurate object examples, and without an accurate detection model, high-quality proposals cannot be discovered [17]. Fig. 1 shows two common types of imprecise pseudo labels: (a) *Part domination*: For some classes, the bounding box surrounds only the most discriminative part, *e.g.*, the head of an animal.

Manuscript received 28 April 2022; revised 9 October 2022; accepted 13 December 2022. Date of publication 29 December 2022; date of current version 9 January 2023. This work was supported in part by the Establishment of Key Laboratory of Shenzhen Science and Technology Innovation Committee under Grant ZDSYS20190902093015527 and in part by the Key Program of Technology Research from Shenzhen Science and Technology Innovation Committee under Grant JSGG20220831104402004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Leyuan Fang. *(Corresponding author: Yong Xu.)*

Zhihao Wu, Chengliang Liu, and Jie Wen are with the Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen 518055, China, and also with the Shenzhen Key Laboratory of Visual Object Detection and Recognition, Shenzhen 518055, China (e-mail: horatio_ng@163.com; liucl1996@163.com; jiewen_pr@126.com).

Yong Xu is with the Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen 518055, China, also with the Shenzhen Key Laboratory of Visual Object Detection and Recognition, Shenzhen 518055, China, and also with the Pengcheng Laboratory, Shenzhen 518055, China (e-mail: laterfall@hit.edu.cn).

Jian Yang is with the PCA Laboratory, and the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@njust.edu.cn).

Xuelong Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China (e-mail: xuelong_li@nwpu.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3231744

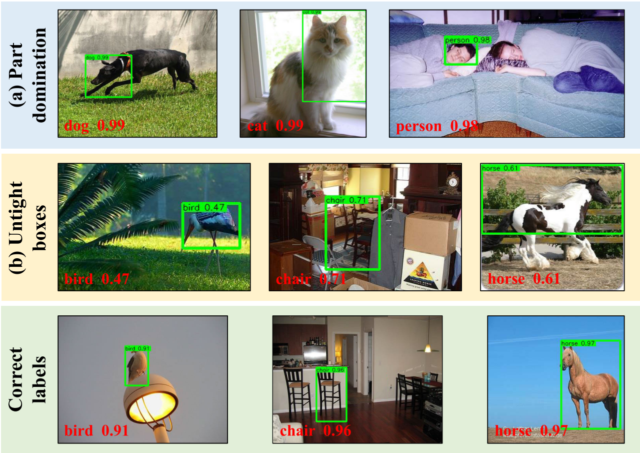

Fig. 1. Typical imprecise pseudo labels produced by online instance classifier refinement (OICR) [16], including (a) part domination and (b) untight boxes. Correct labels can be observed in the third row. The detected category and confidence are plotted on top of the label as well as in the lower left corner of the image. For ease of viewing, only the top-scoring proposal for each class is shown.

(b) *Untight boxes*: The bounding box contains the partial object or some background regions.

To alleviate adverse effects of these two inaccurate labels on detector performance, we focus on selecting high-quality proposals as the supervision for WSOD. Previous works have revealed that the reason behind part domination is that the top layer of a CNN-based classifier focuses more on semantic features rather than location cues [18], [19]. Therefore, an intuitive idea is to incorporate low-level features that contain more location evidence. There have been some works following this line in the fields of WSOD and FSOD. For example, Tang et al. [18] and Zeng et al. [19] exploit low-level information produced by shallow layers and superpixels, respectively, to evaluate the objectness of each proposal. Huang et al. [20] utilize the comprehensive attention map generated by multiple layers to guide feature learning for these layers. However, this approach may be difficult to couple with other WSOD frameworks due to the need to balance many learning branches. Feature pyramid network (FPN) [21] fuses features from multiple layers to improve location representation for fully supervised object detectors. However, it is not adaptable to WSOD since it introduces new layers with learnable parameters, causing the model to converge to an undesirable local minimum [22]. Inspired by [20] and [21], we propose bottom-up aggregated attention (BUAA) to address the problem of part domination, which aims to exploit multi-layer attention maps associated with object localization without additional learning branches or learnable parameters. Different from [20] which forces the attention map from each layer to approximate the comprehensive attention map and [21] which merges multi-level features by convolutional layers, we directly improve the top-level feature representation with the comprehensive map in a spatial attention manner. In this manner, the proposal that covers the complete object or at worst a larger portion of the object can receive a high score. In addition, it does not attach learning branches or learnable parameters. On the one hand, it does not need to set hyperparameters to balance the training of branches, so it can be flexibly plugged into WSOD frameworks. On the other hand, the supervision generation in the beginning of training is highly dependent on the feature extraction ability of the pretrained model, and the BUAA avoids the weakening of this ability caused by adding parameters.

As regards the issue of untight boxes, an intuitive solution is to increase the weights of the loss of instances with correct labels while suppressing the weights of the labels having untight boxes. Therefore, the core question is how to distinguish them. Many previous works [17], [23], [24], [25] follow the assumption that the confidence of a correct box is usually higher than that of an untight one (see the second and third rows in Fig. 1). However, this hypothesis ignores the case where only low-quality proposals are set as pseudo-labels in an image, they may obtain high scores after several instance refinements because they dominate the training of previous classifiers. To solve this problem, we propose to measure the supervision quality using the loss in the instance mining phase. Specifically, we consider that the larger loss in the phase of instance mining, the lower overall quality of the labels in the phase of instance refinement, and vice versa. The reason is as follows: A large loss indicates that the model has low confidence that the image contains positive instances, so even the highest-scoring proposal is probably to be unreliable. Without accurate initial supervision, the pseudo labels generated in instance refinement stages are likely imprecise as well. Based on this assumption, we propose a phase-aware loss (PA-loss) whose modulating factor w.r.t the loss in the instance mining phase is used to reweight the loss in the instance refinement phase. In addition, inspired by the existing methods, we also design a loss weight w.r.t the proposal score, which cooperates with the phase-aware loss to further balance the impact of correct boxes and untight ones on training.

In this work, we integrate the proposed modules into the popular two-phase learning framework OICR [16]. The experimental results on the PASCAL VOC 2007 and 2012 [13] and the MS COCO 2014 [14] benchmarks demonstrate that our method can bring substantial improvement for WSOD and achieve the state-of-the-art performance. Here, we summarize our main contributions:

- We propose bottom-up aggregated attention to alleviate the issue of part domination in WSOD. Compared with previous methods, it can incorporate multi-layer features to compensate for location information without attaching learning branches or learnable parameters.
- We propose a phase-aware loss, which is the first work to measure the quality of pseudo labels via the loss in the instance mining phase, to mitigate the adverse effect of the untight-box problem on training.
- We test the proposed approach on the PASCAL VOC and the MS COCO, yielding the state-of-the-art performance.

II. Related Work

Recent WSOD methods usually follow a two-phase paradigm, including instance mining and instance refinement.

This paradigm suffers from two types of imprecise pseudo labels, *i.e.*, part domination and untight boxes. Many works address the two issues from different respects.

To overcome the challenge of part domination, some research exploits *context information*. Tang et al. [16] assign the same label to the top-scoring positive proposal and its adjacent proposals. Kosugi et al. [26] utilize the classification loss of the context to guide proposal selection. Zhang et al. [27] model the relationship of highly overlapped instances via a graph convolutional network to fuse the features of discriminative portions belonging to the same object. Jia et al. [28] quantify the inclusion relationships between two proposals to discovery complete object. Wu et al. [29] enhance less discriminative features within the neighborhood using local maximum. Some studies *remove discriminative regions*. Gao et al. [30] use two instance mining stages to force the detector to locate the instance with the second highest score, which usually corresponds to the more complete object region. Xu et al. [31] further employ multiple serial mining stages to gradually cover large object regions. Gao et al. [32] adopt multiple parallel mining stages to obtain complementary object parts and fuse them to get the entire object. Ren et al. [33] zero out the pixels belonging to the most discriminative part to force the network to focus on its surrounding object region. Some works *combine with segmentation*. Shen et al. [34] and Li et al. [35] leverage the complementary information of weakly supervised object detection and segmentation tasks. Some efforts exploit *low-level features*. Tang et al. [18] utilize features from early convolutional layers to generate proposals. Zeng et al. [19] combine superpixel evidence and CNN confidence to measure the objectness of proposals. Huang et al. [20] take full advantages of the comprehensive attention map generated by multiple layers to guide their feature learning. In addition, some methods introduce min-entropy [36], feature correlation [37], knowledge transfer [38], [39], continuation optimization [40] and other supervisory signals [41].

Among all approaches mentioned above, [20] is most relevant to our BUAA. Both methods exploit aggregated attention from multiple convolutional layers to boost location representation. However, [20] is a loss regularization method that makes the attention map from each layer approximate the aggregated map, while BUAA is a feature reweighting method which regards the aggregated map as a spatial attention module to refine the top-layer features. This manner does not add learning branches, so it can be easily integrated into WSOD frameworks. Besides, BUAA is also related to FPN [21] and its variants [42], [43] which construct top-down or bottom-up connections to merge features from multiple layers. Compared with them, BUAA does not adopt the connections with convolutional layers. This is because WSOD relies heavily on the initial extraction capability of the model due to the lack of precise supervision. Therefore, fusion layers with learnable parameters may adversely affect the availability of the pretrained model, causing the training to deviate from the desirable solutions.

To resolve the problem of untight boxes, some research focuses on *generating high-quality proposals*. Cheng et al. [24] introduce gradient-weighted class activation mapping in proposal generation to obtain more proposals which have high intersection-over-union (IOU) with ground truth. Some studies explore how to *select reliable proposals*. Many of them [23], [24], [25], [44] increase the weights of labels having high confidence based on the assumption that scores of correct boxes are usually higher than those of untight ones. In addition, Zhang et al. [17] measure the difficulty of the image by the score distribution and feed images with increasing difficulty in the detector during training. Shen et al. [45] introduce the global shape information of objects as location cues.

The proposed PA-loss also follows the idea of emphasizing correct labels and suppressing the labels with untight boxes. Different from the abovementioned approaches, we propose a novel perspective to distinguish between correct boxes and untight ones. To be specific, we consider that a large loss in the first phase indicates the supervision with untight boxes in the second phase. Additionally, the PA-loss is also related to instance sampling, such as online hard example mining [46] and focal loss [7]. These methods adjust the gradient contributions of instances based on their attributes to exploit the potential of detectors. However, different from [7], [46] which quantify the attribute according to the instance error, our PA-loss is based on the loss in the instance mining phase. Also, we present a reweighting scheme based on the proposal score, which shares the same idea with [23] but is more effective.

III. Method

As we mentioned above, this work aims to select high-quality instances as the supervision for WSOD to address the problems of part domination and untight boxes. To this end, we implement our approach on the widely used two-phase framework OICR [16] by introducing two modules, *i.e.*, bottom-up aggregated attention and phase-aware loss. In this section, we first introduce the pipeline of the framework. Then, we detail the proposed modules.

A. Overview Pipeline

As shown in Fig. 2, given an image $\mathcal{I}$, generated proposals $\mathcal{P} = \{p_r\}_{r=1}^{R}$, where R denotes the number of proposals, and image label $\mathbf{x} = [x_1, x_2, \ldots, x_C]^T$, where $x_{c'} = 1$ indicates that at least one object is from the c'-th class while $x_{c'} = 0$ denotes that there are no objects from the c'-th class, and C is the total number of categories. The image $\mathcal{I}$ and proposals $\mathcal{P}$ are first input into the backbone with bottom-up aggregated attention to produce image feature maps. Next, a Spatial Pyramid Pooling (SPP) layer and two fully connected (FC) layers are performed to extract feature vectors for the proposals. After that, the vectors are fed into $\{K + 1\}$ branches: one for the instance mining phase and the others for the instance refinement phase.

In the phase of instance mining, two parallel FC layers are performed to generate two matrices $\mathbf{X}^{\text{cls}}, \mathbf{X}^{\text{det}} \in \mathbb{R}^{C\times R}$. Then, the two matrices are passed through two softmax layers along two dimensions: $\left[\sigma\left(\mathbf{X}^{\text{cls}}\right)\right]_{c'r'} = e^{x^{\text{cls}}_{c'r'}} \Big/ \sum_{c=1}^{C} e^{x^{\text{cls}}_{cr'}}$ and

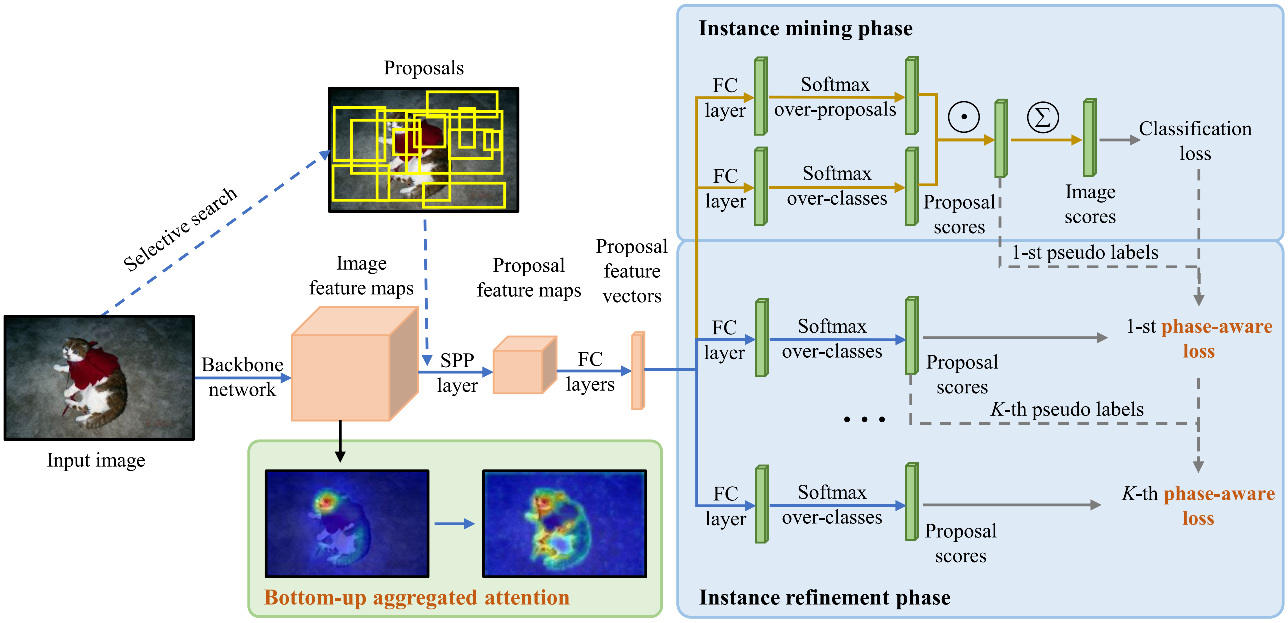

Fig. 2. The pipeline of the two-phase framework which integrates two novel modules: bottom-up aggregated attention and phase-aware loss. Bottom-up aggregated attention aims to improve location representation for the backbone network. Phase-aware loss aims to increase the weights of correct boxes and decrease the weights of untight ones. All arrows indicate the training procedure, solid ones have back-propagation computations, and blue ones show the testing process.

$[\sigma(\mathbf{X}^{\text{det}})]_{c'r'} = e^{x^{\text{det}}_{c'r'}} \big/ \sum_{r=1}^{R} e^{x^{\text{det}}_{c'r}}$. The former indicates the probability that the r'-th proposal is from the c'-th class. The latter represents the normalized contribution of the r'-th proposal to the image containing the object from the c'-th category. The proposal scores are generated by element-wise product: $\varphi^0 = \sigma(\mathbf{X}^{\text{cls}}) \odot \sigma(\mathbf{X}^{\text{det}})$, and then φ^0 is summed along the dimension of the proposal to obtain the image score of each class: $\phi_c = \sum_{r=1}^{R} \varphi^0_{cr}$. Finally, a multi-class cross entropy loss is applied to train this phase: $L^0_{CE} = -\sum_{c=1}^{C} x_c \log \phi_c + (1 - x_c) \log (1 - \phi_c)$.

In the phase of instance refinement, proposal scores $\varphi^{k'} \in \mathbb{R}^{(C+1)\times R}$ are generated in a similar way as in the first phase, where the $\{C+1\}$-th dimension corresponds to background. The corresponding pseudo labels $\mathcal{Y}^{k'} = \{[y^{k'}_{1r}, \dots, y^{k'}_{(C+1)r}]^T\}_{r=1}^{R}$ are provided by the proposal scores from the preceding stage (*i.e.*, $\varphi^{k'-1}$), where $y^{k'}_{c'r'} = 1$ represents that the r'-th proposal belongs to the c'-th class in the k'-th stage. Then, each stage is trained under the guidance of our proposed phase-aware loss, denoted as $L^{k'}_{PA}$ (see Eq. (4) and Eq. (5)).

Finally, the above loss functions are combined to jointly train the framework: $L = L^0_{CE} + \sum_{k=1}^{K} L^k_{PA}$. We summarize the overall training procedure in Algorithm 1.

B. Bottom-Up Aggregated Attention

Our goal is to fully exploit the spatial cues contained in multiple convolutional layers by means of spatial attention mechanism. To this end, we first review the popular spatial attention structure [48], [49], [50]. Given a feature map $\mathcal{F} \in \mathbb{R}^{H\times W\times D}$, where H, W and D represent the height, the width and the depth, respectively. The attention module takes $\mathcal{F}$ as input and generates a normalized attention map $\mathbf{A} \in \mathbb{R}^{H\times W}$ by some convolutional layers. After that, $\mathbf{A}$ is multiplied to $\mathcal{F}$ to obtain refined feature maps $\hat{\mathcal{F}}$ as input to the next layer. However, the convolutional layers contain learnable parameters which may cause the training to deviate from desirable solutions [22]. Therefore, one challenge is how to generate $\mathbf{A}$ without additional learnable parameters. Another challenge is how to aggregate multi-scale attention maps. For the same reason, connections having convolutional operations are not adaptable.

Algorithm 1 The Overall Training Procedure (Each Iteration)

```
Input: An image $\mathcal{I}$, generated proposals $\mathcal{P}$ and the image-category label $\mathbf{x}$; refinement times $K$.
Output: Updated network parameters.
Input the image $\mathcal{I}$ and the proposals $\mathcal{P}$ into the network with the bottom-up aggregated attention module to produce proposal scores $\varphi^k$, where $k \in \{0, 1, \dots, K\}$.
Compute the multi-class cross entropy loss $L^0_{CE}$.
for $k = 1$ to $K$ do
    Produce pseudo labels $\mathcal{Y}^k$ based on the proposal scores $\varphi^{k-1}$.
    Compute the phase-aware loss $L^k_{PA}$.
end for
Optimize $L^0_{CE} + \sum_{k=1}^{K} L^k_{PA}$.
```

To overcome the above challenges, we perform spatial min-max normalization on the resized feature maps to produce attention maps and then fuse them in a bottom-up fashion via an element-wise max operation. Formally, the network contains Q convolutional blocks $\{B_1, \dots B_Q\}$, and the corresponding feature maps are denoted as $\{\mathcal{F}^{B_1}, \dots, \mathcal{F}^{B_Q}\}$, where $\mathcal{F}^{B_Q} \in \mathbb{R}^{H\times W\times D}$ represents the top-layer feature map.

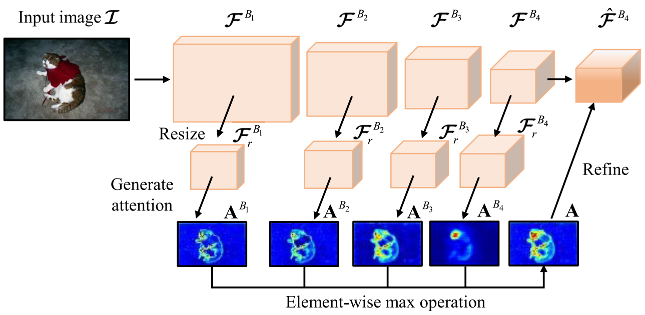

Fig. 3. The example of applying BUAA on VGG-16 [47].

Algorithm 2 The Process of Bottom-up Aggregated Attention

```
Input: Feature maps $\{\mathcal{F}^{B_1}, \ldots, \mathcal{F}^{B_Q}\}$.
Output: The refined top-level feature map $\hat{\mathcal{F}}^{B_Q}$.
1: Resize feature maps $\{\mathcal{F}^{B_1}, \ldots, \mathcal{F}^{B_Q}\}$ to the same height and width as $\mathcal{F}^{B_Q}$ by MaxPool.
2: Generate attention maps $\{\mathbf{A}^{B_1}, \ldots, \mathbf{A}^{B_Q}\}$ using Eq. (1).
3: Compute the aggregated attention map $\mathbf{A}$ using Eq. (2).
4: Refine the top-level feature map $\mathcal{F}^{B_Q}$ using Eq. (3).
```

As illustrated in Fig. 3, we first resize $\{\mathcal{F}^{B_1}, \ldots, \mathcal{F}^{B_Q}\}$ to the same height and width as $\mathcal{F}^{B_Q}$ by MaxPool, denoted as $\left\{\mathcal{F}_r^{B_1}, \ldots, \mathcal{F}_r^{B_Q}\right\}$. Next, the attention map $\mathbf{A}^{B_q} \in \mathbb{R}^{H \times W}$ for each block is computed by channel-wise average pooling and spatial min-max normalization:

$$\mathbf{A}^{B_q} = \frac{\mathbf{F}^{B_q} - \min\limits_{\substack{1 \le x \le W, \\ 1 \le y \le H}} \mathbf{F}^{B_q}(x, y)}{\max\limits_{\substack{1 \le x \le W, \\ 1 \le y \le H}} \mathbf{F}^{B_q}(x, y) - \min\limits_{\substack{1 \le x \le W, \\ 1 \le y \le H}} \mathbf{F}^{B_q}(x, y)}, \tag{1}$$

where $\mathbf{F}^{B_q} = \frac{1}{D} \sum_{d=1}^{D} \left(\mathcal{F}_r^{B_q}\right)_d$. Then, we aggregate the attention maps as follows:

$$\mathbf{A} = \max\left(\mathbf{A}^{B_1}, \ldots, \mathbf{A}^{B_Q}\right), \tag{2}$$

where $\max(\cdot)$ is the element-wise max function, and $\mathbf{A}$ denotes the comprehensive attention map. Finally, we refine the top-level feature map $\mathcal{F}^{B_Q}$ by $\mathbf{A}$:

$$\hat{\mathcal{F}}^{B_Q} = (\mathbf{Y} + \mathbf{A}) \otimes \mathcal{F}^{B_Q}, \tag{3}$$

where $\mathbf{Y} \in \mathbb{R}^{H \times W}$ is a matrix whose elements are all 1, $\otimes$ denotes element-wise multiplication, and $\hat{\mathcal{F}}^{B_Q}$ represents the updated top-level feature map after refinement. The overall process is summarized in Algorithm 2. It is worth noting that min-max normalization+element-wise max is not the only choice to produce aggregated attention. We also explore other approaches such as generating attention maps with the element-wise Sigmoid function and aggregating the maps using an element-wise average operation. These methods also achieve performance gains, but not as much as min-max normalization+element-wise max.

C. Phase-Aware Loss

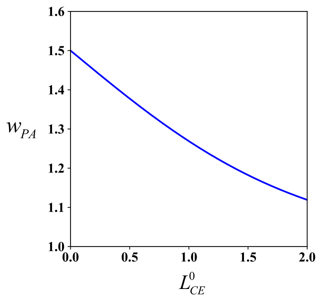

Fig. 4. The proposed w_{PA} w.r.t L_{CE}^0.

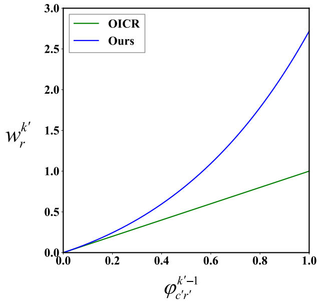

Fig. 5. The proposed $w_r^{k'}$ w.r.t $\varphi_{c'r'}^{k'-1}$.

We aim to boost the weights of the loss of instances having correct labels and suppress the weights of instances with untight boxes. The core question is how to identify them since the ground truth of the bounding box is unknown. In this work, we propose that the loss in the phase of instance mining is an important indicator to measure the supervision quality. Specifically, a large loss indicates that the current model has low confidence that the image contains positive instances. Therefore, the generated supervision is probably unreliable. On the one hand, without accurate initial supervision, the proposals mined in the instance refinement phase are also likely to be imprecise. On the other hand, these inaccurate examples dominate the training, leading them to gradually obtain high scores in refinement. Thus the confidence-based strategies will mistake them for high-quality proposals, causing the training to deviate from the correct direction. To sum up, a large loss in the mining phase usually implies overall low-quality pseudo labels in the refinement phase, and vice versa.

Based on the above analysis, we propose to add a modulating factor $w_{PA} = 1 + \sigma\left(-L_{CE}^0\right)$, where $\sigma(\cdot)$ is the sigmoid function, to the weighted softmax loss function used in OICR [16]. We define the phase-aware loss in the k'-th instance refinement stage as:

$$L_{PA}^{k'} = -\frac{1}{R} \sum_{r=1}^{R} \sum_{c=1}^{C+1} w_{PA} w_r^{k'} y_{cr}^{k'} \log \varphi_{cr}^{k'}, \tag{4}$$

where $w_r^{k'} = \varphi_{c'r'}^{k'-1}$, r' corresponds to the top-scoring proposal (*i.e.*, the core proposal) from the c'-th class in the $\{k' - 1\}$-th stage, and r corresponds to the highest-scoring proposal and its adjacent ones. The modulating factor w_{PA} is visualized in Fig. 4. We can observe that when L_{CE}^0 increases, w_{PA} and $L_{PA}^{k'}$ gradually decrease. Thus the contribution of the corresponding instance to training is relatively reduced.

TABLE I

ABLATION EXPERIMENTS OF THE BUAA HAVING DIFFERENT CONFIGURATIONS ON THE PASCAL VOC 2007 DATASET

Case	Method of generating attention	Method of aggregating attention	Aggregated attention maps	mAP	CorLoc
(a)	-	-	-	48.3	64.4
(b)	Min-max normalization	Element-wise max	A^{B_3}, A^{B_4}	48.8	64.7
(c)			$A^{B_2}, A^{B_3}, A^{B_4}$	49.8	65.6
(d)			$A^{B_1}, A^{B_2}, A^{B_3}, A^{B_4}$	**50.3**	**66.1**
(e)	Min-max normalization	Element-wise average	$A^{B_1}, A^{B_2}, A^{B_3}, A^{B_4}$	48.7	65.2
(f)	Element-wise Sigmoid	Element-wise max		49.0	64.7
(g)	Element-wise Sigmoid	Element-wise average		48.4	64.6

Following OICR, we also design a loss weight w.r.t the proposal score to better measure the quality of pseudo labels:

$$w_r^{k'} = \varphi_{c'r'}^{k'-1} \times e^{\varphi_{c'r'}^{k'-1}}. \qquad (5)$$

As shown in Fig. 5, compared with OICR, we assign more weights to labels with high confidence. The explanation is as follows. For the r''-th proposal in the k'-th stage, the weighted softmax loss function is: $L_{r''}^{k'} = -\sum_{c=1}^{C+1} w_{r''}^{k'} y_{cr''}^{k'} \log \varphi_{cr''}^{k'}$. Suppose $y_{c'r''}^{k'} = 1$ and $y_{c''r''}^{k'} = 0, c'' \neq c'$, and let $\theta_{r''}^{k'}$ be the output of the last FC layer. The gradient of $L_{r''}^{k'}$ is: $\partial L_{r''}^{k'} \Big/ \partial \theta_{r''}^{k'} = w_{r''}^{k'} \left(\varphi_{c'r''}^{k'} - 1 \right)$. If the confidence of the core proposal $\varphi_{c'r'}^{k'-1}$ in the preceding stage is high, the proposal score $\varphi_{c'r''}^{k'}$ in the current stage is probably high since the model can easily discover this object. At this time, $\left| \varphi_{c'r''}^{k'} - 1 \right|$ is small. Conversely, if $\varphi_{c'r'}^{k'-1}$ is low, $\varphi_{c'r''}^{k'}$ is likely low due to the lack of accurate supervision. Accordingly, $\left| \varphi_{c'r''}^{k'} - 1 \right|$ is large. For simplicity, we suppose that $\varphi_{c'r'}^{k'-1}$ and $\varphi_{c'r''}^{k'}$ are equal. If we set $w_{r''}^{k'}$ to $\varphi_{c'r'}^{k'-1}$ as in OICR, $\varphi_{c'r'}^{k'-1} = a$ (*e.g.*, 0.9) and $\varphi_{c'r'}^{k'-1} = 1 - a$ (*e.g.*, 0.1) will contribute the same to training, *i.e.*, their gradients are equal. Obviously, this is the opposite of our purpose to boost the gradient contribution of the instance with the high-scoring label. Therefore, we assign much more weights to the labels with high scores against the original assignment of gradients by softmax loss. In this work, we cooperate the designed loss weight with the PA-loss to further balance the impact of correct boxes and untight ones on training.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We evaluate our approach on the PASCAL VOC 2007 and 2012 [13] and the MS COCO 2014 [14] benchmarks. The VOC 2007 contains 9, 963 images from 20 categories, 5, 011 of which are for training and validation, and the rest for testing. The VOC 2012 contains 22, 531 images, including 11, 540 images for training and validation, and 10, 991 images for testing. In the MS COCO 2014, the *train* set consists of 82, 783 images, and the *val* set has 40, 504 images.

For the PASCAL VOC, we adopt mean average precision (mAP) on the *test* sets and correct localization (CorLoc) on the *trainval* sets for evaluation. mAP is used to measure the object detection performance, which follows the PASCAL VOC protocol of IOU > 0.5 between predicted boxes and the ground truth ones. CorLoc denotes the percentage of training images where the top-scoring proposals have more than 0.5 IOU with the ground truth boxes. For the MS COCO, we use average precision (AP) (averaged over IOU thresholds [0.5 : 0.05 : 0.95]) and AP_{50} (AP for IOU threshold 0.5) on the *val* set.

B. Implementation Details

Our method is built on the OICR framework [16]. We use VGG-16 [47] pretrained on ImageNet [51] as the backbone. Selective search [52] and multiscale combinatorial grouping [53] are performed to provide initial proposals on the PASCAL VOC and the MS COCO, respectively. The added FC layers are initialized by Gaussian distribution with mean 0, standard deviation 0.01 and bias 0. K is set to 3 for the instance refinement phase. We also apply data augmentation: During training, each image is randomly flipped horizontally, and its shorter side is randomly resized into five scales $\{480, 576, 688, 864, 1200\}$ and the longer side is no more than 2000. During testing, all ten augmented images are used. The framework is optimized by SGD with weight decay $5e^{-4}$ and momentum 0.9. The batch size is 4, and the maximum iteration numbers are 37, 500, 55, 000 and 200, 000 for the VOC 2007, the VOC 2012 and the MS COCO, respectively. For the VOC 2007 and 2012, the initial learning rate is $5e^{-4}$ for the first 27, 500 and 47, 500 iterations and then is decreased by a factor 0.1. For the MS COCO, the learning rate is $1e^{-3}$ for the first 150, 000 iterations and $1e^{-4}$ for the following iterations. The threshold for non-maximum suppression is set to 0.3, which is used to filter out highly overlapping bounding boxes.

C. Ablation Experiments

Our method introduces two modules to OICR [16], including bottom-up aggregated attention and phase-aware loss. To evaluate the contributions of the two modules, we conduct ablation experiments based on OICR. For all the experiments, we adopt the same hyperparameters, except for the specific changes to the module being evaluated. The experiments are implemented on the PASCAL VOC 2007 benchmark.

1) Influence of Bottom-Up Aggregated Attention: We first discuss the influence of the BUAA having different configurations on the detector. In Table I, case (a) corresponds to OICR. Cases (b) to (d) aggregate attention maps from different layers via min-max normalization+element-wise max. Cases (e) to (g) use different methods to aggregate attention maps from all levels. Cases (b) to (g) all bring performance improvement, which fully demonstrate the effectiveness of our

TABLE II

THE CLASSES THAT ACHIEVE ABOVE-AVERAGE AP_{50} GAINS ON THE PASCAL VOC 2007 DATASET

Method	bird	boat	cat	dog	horse	person	sheep	mAP
OICR	51.3	19.7	39.2	37.4	48.8	13.9	45.2	48.3
OICR with BUAA	56.2 (**+4.9**)	26.3 (**+6.6**)	55.6 (**+16.4**)	59.9 (**+22.5**)	57.5 (**+8.7**)	16.9 (**+3.0**)	49.1 (**+3.9**)	50.3 (**+2.0**)

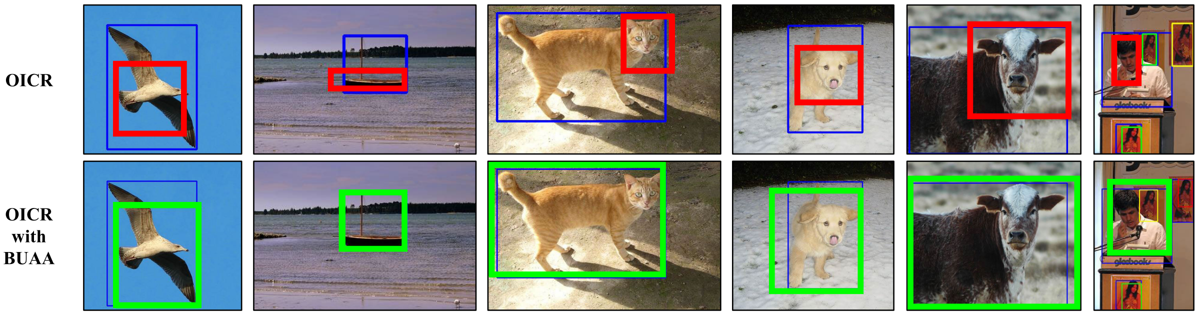

Fig. 6. Qualitative comparisons between OICR and OICR with BUAA on the PASCAL VOC 2007. The blue rectangle represents the ground truth which has no less than one result with IOU > 0, and the yellow one represents the ground truth having no result intersection. The success case (IOU > 0.5) is indicated by the green rectangle, and the failure case (IOU < 0.5) is indicated by the red one. The objects for comparison are highlighted with thick rectangles.

strategy to refine the top-level feature representation by multi-layer features in a spatial attention manner. Among all cases, case (d) obtains the best results and brings 2.0% mAP gain and 1.7% CorLoc gain compared to OICR. This is because min-max normalization can better localize discriminative features by synthesizing global information, and element-wise max can make full use of salient features extracted by each level. All other experiments apply the configuration in case (d).

We further investigate which classes the performance gains mainly come from. Specifically, we compare the AP_{50} results for each class of the OICR with and without BUAA, and then show the classes that achieve above-average gains in Table II. We can observe that the animals and the boat get the significant performance boost, and these classes are also prone to getting stuck in the most discriminative parts, *e.g.*, heads of animals and cabins of boats. Moreover, we compare the visualized detection results of OICR and the OICR having BUAA. From Fig. 6, we can see that cooperating with BUAA, the detector can discover more complete object regions. The quantitative and qualitative results demonstrate that the proposed BUAA can effectively alleviate the problem of part domination.

We also construct an FPN-like connection architecture to leverage multi-layer features. To be specific, we first perform a MaxPool operation and a 1×1 convolutional layer on the lowest-level feature map to ensure that it has the same size as its upper map, and then merge them by element-wise addition and a 3×3 convolution. After that, we repeat the above process on the merged map until the highest-level features are fused. This approach only achieves 3.6% mAP and 25.6% CorLoc, which are much lower than the results of BUAA (50.3% mAP and 66.1% CorLoc). This is because the additional convolutional layers with learnable parameters make the pretrained model lose the initial feature extraction ability, causing the model to fail to converge to desirable solutions. Therefore, our designed connection method without learnable parameters is necessary and effective.

In addition, we compare our BUAA with layer-wise comprehensive attention self-distillation (LW-CASD) [20]. Built upon our baseline, LW-CASD receives 49.9% mAP and 65.8% CorLoc, which are 0.4% mAP and 0.3% CorLoc lower than BUAA. Furthermore, BUAA does not add new learning branches. These illustrate the superiority of our strategy.

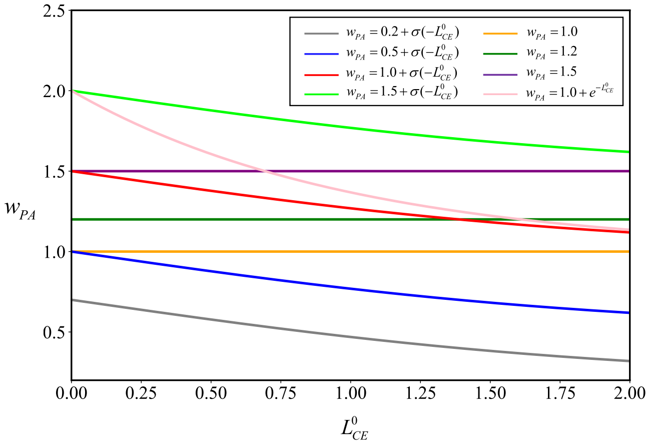

Fig. 7. The visualization of w_{PA} with different settings.

TABLE III

ABLATION EXPERIMENTS OF THE PA-LOSS WITH DIFFERENT CONFIGURATIONS ON THE PASCAL VOC 2007 DATASET

Case	w_{PA}	$w_r^{k'}$	mAP	CorLoc
(a)	1.0	$\varphi_{c'r'}^{k'-1}$	48.3	64.4
(b)	1.2		49.3	65.0
(c)	1.5		50.1	65.7
(d)	$0.2+\sigma\left(-L_{CE}^0\right)$		48.6	65.3
(e)	$0.5+\sigma\left(-L_{CE}^0\right)$		49.9	66.0
(f)	$1.0+\sigma\left(-L_{CE}^0\right)$		**51.5**	**66.6**
(g)	$1.5+\sigma\left(-L_{CE}^0\right)$		51.3	66.5
(h)	$1.0+e^{-L_{CE}^0}$		51.0	65.8
(i)	1.0	$\varphi_{c'r'}^{k'-1} \times e^{\varphi_{c'r'}^{k'-1}}$	52.8	67.9
(j)	$1.0+\sigma\left(-L_{CE}^0\right)$		**53.4**	**68.0**

2) Influence of Phase-Aware Loss: We investigate the influence of the PA-loss with different configurations. As we stated in Section III-C, the PA-loss has two tunable factors,

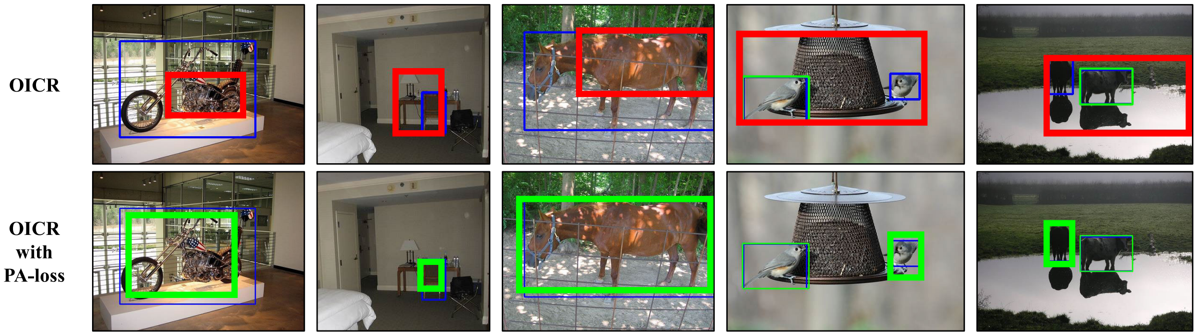

Fig. 8. Qualitative comparisons between OICR and OICR with PA-loss on the PASCAL VOC 2007. The blue rectangle represents the ground truth. The success case (IOU > 0.5) is indicated by the green rectangle, and the failure case (IOU < 0.5) is indicated by the red one. The objects for comparison are highlighted with thick rectangles.

TABLE IV

Ablation Experiments of BUAA and PA-Loss on the PASCAL VOC 2007 Dataset

BUAA	PA-loss	mAP	CorLoc
-	-	48.3	64.4
✓	-	50.3	66.1
-	✓	53.4	68.0
✓	✓	**54.3**	**70.0**

i.e., w_{PA} and $w_r^{k'}$. The two factors are respectively based on the following assumptions: A low loss in the instance mining phase indicates high-quality supervision, and a high-confidence proposal corresponds to a correct pseudo label. To evaluate the effectiveness of w_{PA}, we set $w_r^{k'} = \varphi_{c'r'}^{k'-1}$ and change the setting of w_{PA} (see Fig. 7 and cases (a) to (h) in Table III), where $w_{PA} = 1.0$ (case (a)) corresponds to the setting in OICR. Cases (b) and (c) set w_{PA} to a fixed value and naively increase the weight of instance refinement phase. Cases (d) and (e) make w_{PA} decrease as L_{CE}^0 increases and suppress the weight of the refinement phase. Cases (f) to (h) also let w_{PA} decrease as L_{CE}^0 increases but up-weight the refinement phase. It can be seen that cases (b) to (e) outperform the baseline but are inferior to cases (f) to (h). This observation demonstrates that the improvement brought by our strategy of adjusting the weight based on the loss in the phase of instance mining is without any bells and whistles. In other words, this verifies our viewpoint that a small loss in the mining phase usually corresponds to pseudo labels with tight boxes in the refinement phase. Among all settings, the PA-loss with $w_{PA} = 1 + \sigma\left(-L_{CE}^0\right)$ achieves the best results, *i.e.*, 51.5% mAP and 66.6% CorLoc, outperforming the baseline by 3.2% mAP and 2.2% CorLoc.

Next, we evaluate the effectiveness of $w_r^{k'}$. We first set $w_r^{k'}$ in the weighted softmax loss function of OICR to $\varphi_{c'r'}^{k'-1} \times e^{\varphi_{c'r'}^{k'-1}}$. In this way, we get an mAP of 52.8% and a CorLoc of 67.9%, which are 4.5% mAP and 3.5% CorLoc higher than the baseline. We further cooperate it with $w_{PA} = 1 + \sigma\left(-L_{CE}^0\right)$. This full version of PA-loss can boost mAP by 5.1% (from 48.3% to 53.4%) and CorLoc by 3.6% (from 64.4% to 68.0%). In Fig. 8, we show qualitative comparisons between the vanilla OICR and the OICR with the full PA-loss.

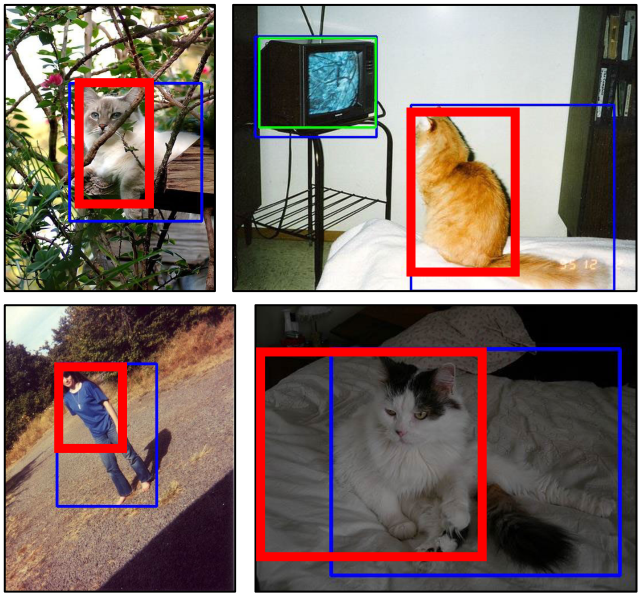

Fig. 9. Some failure detection results of our full model on the PASCAL VOC 2007. The blue rectangle is the ground truth. The success case (IOU > 0.5) is represented by the green rectangle, and the failure case (IOU < 0.5) is represented by the thick red one.

The results demonstrate that the PA-loss can effectively alleviate the problem of untight bounding boxes by focusing more on the labels corresponding to the low loss in the instance mining phase and the proposals having high scores. Furthermore, we perform the PA-loss on the OICR with BUAA. As shown in Table IV, it can significantly improve mAP from 50.3% to 54.3% and CorLoc from 66.1% to 70.0%.

D. Comparison With Other Methods

We compare the full version of our approach (BUAA-PAL for short) with other state-of-the-art methods that use VGG-16 as the backbone. To further demonstrate the validity of our approach, we also adopt the OICR with a bounding-box regression branch (Reg. in the tables) [60] as a stronger baseline.

1) Results on the PASCAL VOC 2007: Table V and VI show the results of our methods and competing approaches on the VOC 2007 dataset. A total of 25 state-of-the-art WSOD

TABLE V

RESULTS (MAP) OF THE STATE-OF-THE-ART WSOD METHODS ON THE PASCAL VOC 2007 DATASET

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
PCL [23]	54.4	69.0	39.3	19.2	15.7	62.9	64.4	30.0	25.1	52.5	44.4	19.6	39.3	67.7	17.8	22.9	46.6	57.5	58.6	63.0	43.5
TS^{2}C [54]	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
WSRPN [18]	57.9	70.5	37.8	5.7	21.0	66.1	69.2	59.4	3.4	57.1	**57.3**	35.2	64.2	68.6	32.8	28.6	50.8	49.5	41.1	30.0	45.3
MELM [36]	55.6	66.9	34.2	29.1	16.4	68.8	68.1	43.0	25.0	65.6	45.3	53.2	49.6	68.6	2.0	25.4	52.5	56.8	62.1	57.1	47.3
CSC [45]	51.4	62.0	35.2	18.7	27.9	66.7	53.5	51.4	16.2	43.6	43.0	46.7	20.0	58.4	31.1	23.8	43.6	48.8	65.4	53.5	43.0
C-MIL [40]	62.5	58.4	49.5	32.1	19.8	70.5	66.1	63.4	20.0	60.5	52.9	53.5	57.4	68.9	8.4	24.6	51.8	58.7	66.7	63.5	50.5
WS-JDS [34]	52.0	64.5	45.5	26.7	27.9	60.5	47.8	59.7	13.0	50.4	46.4	56.3	49.6	60.7	25.4	28.2	50.0	51.4	66.5	29.7	45.6
SDCN [35]	59.4	71.5	38.9	32.2	21.5	67.7	64.5	68.9	20.4	49.2	47.6	60.9	55.9	67.4	31.2	22.9	45.0	53.2	60.9	54.4	50.2
C-MIDN [30]	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
Pred Net [55]	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
WSOD^2 [19]	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
OIM [44]	55.6	67.0	45.8	27.9	21.1	69.0	68.3	70.5	21.3	60.2	40.3	54.5	56.5	70.1	12.5	25.0	52.9	55.2	65.0	63.7	50.1
OCRepr-OICR [56]	60.4	64.6	44.7	23.5	17.6	65.9	60.8	67.3	24.3	48.5	39.2	49.1	52.8	62.5	11.4	21.3	43.3	51.3	58.8	58.8	46.3
OCRepr-OICR+Reg. [56]	59.4	66.4	45.8	21.5	22.1	70.1	67.3	66.1	24.2	58.8	48.5	60.5	62.4	66.7	17.9	26.0	47.5	57.5	60.5	63.5	50.6
PG-PS [24]	63.0	64.4	50.1	27.5	17.1	70.6	66.0	71.1	25.8	55.9	43.2	62.7	65.9	64.1	10.2	22.5	48.1	53.8	**72.2**	67.4	51.1
GAM+Reg. [57]	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
PSLR [27]	62.2	61.1	51.1	33.8	18.0	66.7	66.5	65.0	18.5	59.4	44.8	60.9	65.6	66.9	24.7	26.0	51.0	53.2	66.0	62.2	51.2
SLV [58]	65.6	71.4	49.0	37.1	24.6	69.6	70.3	70.6	30.8	63.1	36.0	61.4	65.3	68.4	12.4	**29.9**	52.4	60.0	67.6	64.5	53.5
MIST [33]	**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
P-MIDN+MGSC [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	53.9
GradingNet-C-MIL [28]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54.3
IM-CFB [59]	64.1	74.6	44.7	29.4	26.9	73.3	72.0	71.2	28.1	66.7	48.1	63.8	55.5	68.3	17.8	27.7	54.4	62.7	70.5	66.6	54.3
D-MIL [32]	60.4	71.3	51.1	25.4	23.8	70.4	70.3	71.9	25.2	63.4	42.6	**67.1**	57.7	70.1	15.5	26.6	58.7	63.3	66.9	**67.6**	53.5
OICR (baseline) [16]	58.9	69.8	51.3	19.7	25.8	71.0	71.4	39.2	22.0	58.7	47.0	37.4	48.8	72.4	13.9	26.9	45.2	55.0	65.2	66.8	48.3
OICR+Reg. (baseline) [16]	59.0	71.9	43.8	34.2	27.0	73.8	69.9	72.8	29.1	70.9	49.3	58.2	59.2	70.4	10.6	24.7	53.5	63.1	64.2	40.7	52.3
BUAA-PAL-OICR	67.3	78.2	55.5	31.0	22.0	72.9	74.0	74.3	29.8	64.6	51.3	65.4	60.3	72.1	16.8	27.3	54.1	**64.4**	69.9	34.7	54.3
BUAA-PAL-OICR+Reg.	66.1	**80.1**	41.3	30.1	28.5	**75.3**	72.0	**76.2**	**33.5**	69.7	48.6	62.1	60.2	73.1	16.7	26.8	54.1	60.4	70.8	63.3	**55.4**

TABLE VI

RESULTS (CORLOC) OF THE STATE-OF-THE-ART WSOD METHODS ON THE PASCAL VOC 2007 DATASET

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	CorLoc
PCL [23]	79.6	85.5	62.2	47.9	37.0	83.8	83.4	43.0	38.3	80.1	50.6	30.9	57.8	90.8	27.0	58.2	75.3	68.5	75.7	78.9	62.7
TS^{2}C [54]	84.2	74.1	61.3	52.1	32.1	76.7	82.9	66.6	42.3	70.6	39.5	57.0	61.2	88.4	9.3	54.6	72.2	60.0	65.0	70.3	61.0
WSRPN [18]	77.5	81.2	55.3	19.7	44.3	80.2	86.6	69.5	10.1	87.7	**68.4**	52.1	84.4	91.6	**57.4**	**63.4**	77.3	58.1	57.0	53.8	63.8
MELM [36]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.4
CSC [45]	76.1	75.3	61.8	42.0	54.1	74.7	78.8	67.4	32.8	73.1	46.5	59.9	37.6	78.0	56.0	42.5	71.9	67.3	82.4	65.6	62.2
C-MIL [40]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.0
WS-JDS [34]	82.9	74.0	73.4	47.1	**60.9**	80.4	77.5	78.8	18.6	70.0	56.7	67.0	64.5	84.0	47.0	50.1	71.9	57.6	83.3	43.5	64.5
SDCN [35]	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-MIDN [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	68.7
Pred Net [55]	**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
WSOD^2 [19]	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
OIM [44]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.2
OCRepr-OICR [56]	82.9	78.0	67.0	50.0	39.3	79.7	83.2	76.5	37.8	76.7	43.3	67.7	77.2	88.0	12.9	54.2	77.3	62.4	73.8	77.8	65.3
OCRepr-OICR+Reg. [56]	85.4	79.2	65.2	47.9	42.4	84.3	83.3	76.2	37.8	79.5	47.9	71.4	83.7	90.8	25.8	57.9	71.1	64.5	75.3	80.6	67.5
PG-PS [24]	85.4	80.4	69.1	58.0	35.9	82.7	86.7	82.6	45.5	84.9	44.1	**80.2**	84.0	89.2	12.3	55.7	**79.4**	63.4	82.1	82.1	69.2
GAM+Reg. [57]	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
PSLR [27]	86.3	72.9	71.2	59.0	36.3	80.2	84.4	75.6	30.8	83.6	53.2	75.1	82.7	87.1	37.7	54.6	74.2	59.1	79.8	78.9	68.1
SLV [58]	84.6	84.3	73.3	58.5	49.2	80.2	87.0	79.4	46.8	83.6	41.8	79.3	**88.8**	90.4	19.5	59.7	**79.4**	67.7	82.9	**83.2**	71.0
MIST [33]	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
P-MIDN+MGSC [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.8
GradingNet-C-MIL [28]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.0
IM-CFB [59]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	70.7
D-MIL [32]	81.3	82.0	72.7	48.9	42.0	80.2	86.1	78.5	43.9	80.2	42.2	76.5	68.7	91.2	32.7	56.0	81.4	69.6	78.7	79.9	68.7
OICR (baseline) [16]	77.2	81.6	67.9	49.5	45.4	83.3	87.5	51.5	35.0	83.6	50.2	47.9	71.4	91.2	19.4	54.2	77.3	56.5	75.7	81.7	64.4
OICR+Reg. (baseline) [16]	81.1	86.5	61.8	63.4	41.4	82.2	89.7	**85.6**	47.1	83.9	49.6	69.5	82.3	91.2	17.3	53.9	**79.4**	63.5	83.4	62.4	68.8
BUAA-PAL-OICR	84.2	86.7	71.5	52.1	38.2	83.8	87.8	84.3	47.7	80.8	50.6	76.3	79.3	94.0	29.5	61.2	77.3	**70.7**	82.5	60.9	70.0
BUAA-PAL-OICR+Reg.	84.8	**90.4**	61.7	54.2	50.2	**87.8**	86.0	84.1	**50.0**	85.9	49.7	73.1	80.6	**95.0**	28.1	62.9	76.3	60.7	**83.9**	79.1	**71.2**

methods are chosen for comparison. They are PCL [23], TS^{2}C [54], WSRPN [18], MELM [36], CSC [45], C-MIL [40], WS-JDS [34], SDCN [35], C-MIDN [30], Pred Net [55], WSOD^2 [19], OIM [44], OCRepr-OICR [56], OCRepr-OICR+Reg. [56], PG-PS [24], GAM+Reg. [57], PSLR [27], SLV [58], MIST [33], P-MIDN+MGSC [31], GradingNet-C-MIL [28], IM-CFB [59], D-MIL [32], OICR [16] and OICR+Reg. [16]. As we can see, our proposed "BUAA-PAL-OICR+Reg." method outperforms all comparison methods in the metrics of both mAP and CorLoc. Particularly, it improves the mAP of the baseline "OICR+Reg." from 52.3% to 55.4% and the CorLoc from 68.8% to 71.2%. It also outperforms MIST [33] by 0.5% mAP and 2.4% CorLoc, and SLV [58] by 1.1% mAP and 0.2% CorLoc. Built upon the baseline OICR, our full method introduces 6.0% mAP gain (from 48.3% to 54.3%) and 5.6% CorLoc gain (from 64.4% to 70.0%).

2) Results on the PASCAL VOC 2012: Table VII and Table VIII show the detection results of our methods and the state-of-the-art approaches on the PASCAL VOC 2012 dataset. It can be seen that our "BUAA-PAL-OICR" achieves 51.2% mAP and 72.4% CorLoc, outperforming OICR by clear margins of 5.0% mAP and 4.4% CorLoc, respectively. Built upon the OICR with the regression branch, our approach achieves the best performance. Specifically, it yields 53.0% mAP and 74.0% CorLoc, which are 5.0% mAP and 4.5% CorLoc higher than its baseline results. It also outperforms the most competitive "P-MIDN+MGSC" by 0.2% mAP and 0.7% CorLoc.

Furthermore, as can be seen from Table V to VIII, our methods achieve significant performance improvements on the categories prone to part domination (*e.g.*, cat and dog) and untight boxes (*e.g.*, chair and diningtable). This observation

TABLE VII

RESULTS (MAP) OF THE STATE-OF-THE-ART WSOD METHODS ON THE PASCAL VOC 2012 DATASET

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
PCL [23]	58.2	66.0	41.8	24.8	27.2	55.7	55.2	28.5	16.6	51.0	17.5	28.6	49.7	70.5	7.1	25.7	47.5	36.6	44.1	59.2	40.6
TS^{2}C [54]	67.4	57.0	37.7	23.7	15.2	56.9	49.1	64.8	15.1	39.4	19.3	48.4	44.5	67.2	2.1	23.3	35.1	40.2	46.6	45.8	40.0
WSRPN [18]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.8
MELM [36]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	42.4
CSC [45]	54.8	52.2	36.5	18.1	25.4	55.7	39.1	47.2	16.1	39.2	17.9	39.9	34.2	56.1	25.2	20.1	34.6	30.9	56.4	41.4	37.1
C-MIL [40]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.7
WS-JDS [34]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.1
SDCN [35]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	43.5
C-MIDN [30]	72.9	68.9	53.9	25.3	29.7	60.9	56.0	78.3	23.0	57.8	25.7	73.0	63.5	73.7	13.1	28.7	51.5	35.0	56.1	57.5	50.2
Pred Net [55]	73.1	71.4	56.3	**30.8**	28.7	57.6	**62.1**	44.6	23.4	61.7	26.4	44.4	62.7	**80.0**	9.1	24.4	56.8	40.2	52.8	60.8	48.4
WSOD^2 [19]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	47.2
OIM [44]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.3
OCRepr-OICR [56]	69.2	60.2	46.8	25.0	22.4	52.5	52.0	66.7	16.2	49.2	24.8	63.7	59.2	68.7	3.6	23.2	42.6	42.0	40.0	53.9	44.1
OCRepr-OICR+Reg. [56]	70.5	67.1	51.8	27.0	28.3	54.9	57.4	80.6	14.9	56.3	23.3	75.7	66.7	69.4	9.3	24.5	45.0	**50.6**	34.8	57.2	48.3
PG-PS [24]	68.3	60.0	47.4	26.4	20.6	61.5	59.9	**82.1**	23.7	50.4	20.1	**78.8**	52.7	67.7	2.6	21.5	43.8	50.1	**67.2**	60.5	48.3
GAM+Reg. [57]	60.4	68.6	51.4	22.0	25.9	49.4	58.4	62.1	14.5	58.8	24.6	60.4	64.3	70.3	9.4	26.0	47.7	45.5	36.7	55.8	45.6
PSLR [27]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.3
SLV [58]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.2
MIST [33]	**78.3**	73.9	56.5	30.4	**37.4**	64.2	59.3	60.3	26.6	**66.8**	25.0	55.0	61.8	79.3	14.5	**30.3**	**61.5**	40.7	56.4	**63.5**	52.1
P-MIDN+MGSC [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	52.8
GradingNet-C-MIL [28]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	50.5
IM-CFB [59]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.4
D-MIL [32]	69.5	69.5	53.6	23.9	29.2	60.0	58.1	75.0	22.4	60.5	27.4	75.8	64.2	73.0	6.3	23.8	52.7	36.6	51.4	59.1	49.6
OICR (baseline) [16]	75.5	71.0	52.0	28.1	32.3	58.9	59.5	21.3	21.3	61.1	26.5	36.9	57.9	73.4	4.1	27.3	54.5	39.9	59.4	62.1	46.2[§]
OICR+Reg. (baseline) [16]	74.8	**76.4**	57.6	29.2	31.4	59.1	58.6	59.4	**29.4**	56.6	**31.6**	46.8	55.4	78.8	5.1	27.2	55.0	36.4	44.9	47.2	48.0[¶]
BUAA-PAL-OICR	74.4	74.5	58.1	29.5	33.5	58.5	58.5	70.5	25.9	64.9	30.8	59.5	**68.8**	74.0	18.2	29.6	54.0	34.9	51.8	54.7	51.2[†]
BUAA-PAL-OICR+Reg.	74.5	75.4	**60.4**	29.7	31.2	**64.5**	57.9	77.4	25.1	64.5	25.1	72.9	68.4	74.7	**25.7**	28.7	51.3	36.5	59.6	56.6	**53.0**[‡]

[§] http://host.robots.ox.ac.uk:8080/anonymous/PJE1HJ.html [¶] http://host.robots.ox.ac.uk:8080/anonymous/5GNZXN.html
[†] http://host.robots.ox.ac.uk:8080/anonymous/JRWOAN.html [‡] http://host.robots.ox.ac.uk:8080/anonymous/AWW6ZT.html

TABLE VIII

RESULTS (CORLOC) OF THE STATE-OF-THE-ART WSOD METHODS ON THE PASCAL VOC 2012 DATASET

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	CorLoc
PCL [23]	77.2	83.0	62.1	55.0	49.3	83.0	75.8	37.7	43.2	81.6	46.8	42.9	73.3	90.3	21.4	56.7	84.4	55.0	62.9	82.5	63.2
TS^{2}C [54]	79.1	83.9	64.6	50.6	37.8	87.4	74.0	74.1	40.4	80.6	42.6	53.6	66.5	88.8	18.8	54.9	80.4	60.4	70.7	79.3	64.4
WSRPN [18]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.9
MELM [36]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.9
CSC [45]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.4
C-MIL [40]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.4
WS-JDS [34]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.5
SDCN [35]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.9
C-MIDN [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.2
Pred Net [55]	88.8	85.1	68.7	52.3	47.2	**91.0**	**92.1**	64.3	29.4	85.6	54.5	64.9	85.9	89.8	27.5	58.5	81.3	67.6	77.2	79.5	69.5
WSOD^2 [19]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.9
OIM [44]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.1
OCRepr-OICR [56]	87.7	83.5	74.1	53.5	48.0	82.6	76.3	78.2	39.1	85.4	48.8	74.0	85.3	88.1	12.7	57.4	80.4	60.0	61.8	82.9	68.0
OCRepr-OICR+Reg. [56]	88.5	82.1	75.4	55.4	51.0	82.8	78.9	**90.7**	41.8	88.0	46.6	**83.7**	**88.6**	89.4	26.9	58.1	84.4	**72.1**	63.1	85.0	71.6
PG-PS [24]	85.5	81.1	69.2	54.3	37.6	86.7	81.7	84.0	44.6	83.3	45.8	80.2	84.2	87.2	11.5	52.1	78.9	63.9	81.0	80.9	68.7
GAM+Reg. [57]	80.2	83.0	73.1	51.6	48.3	79.8	76.6	70.3	44.1	87.7	50.9	70.3	84.7	92.4	28.5	59.3	83.4	64.6	63.8	81.2	68.7
PSLR [27]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	68.7
SLV [58]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.2
MIST [33]	**91.7**	85.6	71.7	**56.6**	55.6	88.6	77.3	63.4	53.6	**90.0**	51.6	62.6	79.3	94.2	32.7	58.8	**90.5**	57.7	70.9	**85.7**	70.9
P-MIDN+MGSC [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	73.3
GradingNet-C-MIL [28]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.9
IM-CFB [59]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.6
D-MIL [32]	84.5	83.0	71.5	51.9	52.1	89.5	76.7	83.9	51.5	87.7	52.3	82.7	84.5	91.2	19.4	53.0	84.4	50.8	67.8	83.0	70.1
OICR (baseline) [16]	91.2	85.5	67.4	46.9	**57.7**	86.0	89.0	33.4	49.1	87.4	59.8	41.8	73.3	90.8	17.8	64.2	82.5	70.1	82.4	84.7	68.0
OICR+Reg. (baseline) [16]	90.8	**91.1**	72.0	47.8	56.4	86.4	87.0	62.7	**56.2**	81.1	**66.2**	52.8	70.1	**95.2**	19.1	63.6	84.7	66.7	69.4	70.1	69.5
BUAA-PAL-OICR	90.2	88.0	74.0	48.5	56.7	85.7	86.7	76.0	52.0	86.6	62.6	66.8	81.2	94.6	28.2	**66.0**	82.7	65.3	76.8	78.8	72.4
BUAA-PAL-OICR+Reg.	90.2	89.9	**75.8**	48.5	56.2	89.6	84.5	82.9	50.9	86.5	59.6	76.8	82.6	**95.2**	**35.0**	64.3	80.9	66.5	**83.0**	81.9	**74.0**

demonstrates the effectiveness of our proposed BUAA and PA-loss for high-quality proposal selection.

3) Results on the MS COCO 2014: To further demonstrate the effectiveness of our proposed approach, we conduct experiments on the challenging MS COCO dataset. Since only some works have reported their results, here we compare with 14 methods, including PCL [23], MELM [36], CSC [45], WS-JDS [34], C-MIDN [30], WSOD^2 [19], PG-PS [24], PSLR [27], MIST [33], P-MIDN+MGSC [31], GradingNet-C-MIL [28], D-MIL [32], OICR [16] and OICR+Reg. [16]. As shown in Table IX, the "BUAA-PAL-OICR+Reg." method achieves the second highest results of 26.3% AP_{50} and 12.4% AP, which are 1.8% AP_{50} and 1.0% AP higher than the results of its baseline. Without the regression branch, it improves the baseline OICR with 2.7% AP_{50} gain and 1.3% AP gain.

E. Qualitative Results

As stated in Section IV-C, Fig. 6 shows that the model with BUAA can cover the complete object or the large portion of the object rather than the salient part. Fig. 8 displays that the model can alleviate the problem of untight bounding boxes by virtue of the PA-loss. In addition, we also show some failure detection results of our full model in Fig. 9. We can observe that it is difficult for the detector to find the complete object in the following cases: (a) The object is occluded. (b) The bounding box contains a relatively large background region. (c) For different

TABLE IX

COMPARISON WITH THE STATE-OF-THE-ART WSOD METHODS ON THE MS COCO DATASET

Method	AP_{50}	AP
PCL [23]	19.4	8.5
MELM [36]	18.8	-
CSC [45]	20.3	10.2
WS-JDS [34]	20.3	10.5
C-MIDN [30]	21.4	9.6
$WSOD^2$ [19]	22.7	10.8
PG-PS [24]	20.7	-
PSLR [27]	23.6	11.1
MIST [33]	24.3	11.4
P-MIDN+MGSC [31]	**27.4**	**13.1**
GradingNet-C-MIL [28]	25.0	11.6
D-MIL [32]	24.7	11.3
OICR (baseline) [16]	22.9	10.6
OICR+Reg. (baseline) [16]	24.5	11.4
BUAA-PAL-OICR	25.6	11.9
BUAA-PAL-OICR+Reg.	26.3	12.4

individuals belonging to the same class, some object parts have small differences, while other portions are quite different. (d) The object is highly similar to its surrounding background. Thus there is still much room to improve the performance of WSOD.

V. CONCLUSION

In this paper, we proposed a novel approach for weakly supervised object detection (WSOD), which is achieved by attaching two concise but effective modules to the two-phase WSOD framework. One is the bottom-up aggregated attention, which aims to compensate for location information by exploiting multi-layer features without learning branches and learnable parameters. The other one is the phase-aware loss, which increases the weights of the loss of instances with correct labels. Benefiting from the two designs, we can select high-quality instances as the supervision to train the detector. Extensive experiments on the PASCAL VOC and the MS COCO benchmarks validate the superiority of our method.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[2] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[5] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 21–37.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jan. 2017, pp. 2961–2969.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[9] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, "Attention CoupleNet: Fully convolutional attention coupling network for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 113–126, Jan. 2018.

[10] F. Sun, T. Kong, W. Huang, C. Tan, B. Fang, and H. Liu, "Feature pyramid reconfiguration with consistent loss for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5041–5051, May 2019.

[11] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.

[12] Z. Wu, C. Liu, C. Huang, J. Wen, and Y. Xu, "Deep object detection with example attribute based prediction modulation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2020–2024.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[14] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 740–755.

[15] S. Shao et al., "Objects365: A large-scale, high-quality dataset for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8430–8439.

[16] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2843–2851.

[17] X. Zhang, J. Feng, H. Xiong, and Q. Tian, "Zigzag learning for weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4262–4270.

[18] P. Tang et al., "Weakly supervised region proposal network and object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 352–368.

[19] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8292–8300.

[20] Z. Huang, Y. Zou, B. Kumar, and D. Huang, "Comprehensive attention self-distillation for weakly-supervised object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16797–16807.

[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[22] Y. Shen, R. Ji, Z. Chen, Y. Wu, and F. Huang, "UWSOD: Toward fully-supervised-level capacity weakly supervised object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7005–7019.

[23] P. Tang et al., "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.

[24] G. Cheng, J. Yang, D. Gao, L. Guo, and J. Han, "High-quality proposals for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 5794–5804, 2020.

[25] Z. Wu, J. Wen, Y. Xu, J. Yang, and D. Zhang, "Multiple instance detection networks with adaptive instance refinement," *IEEE Trans. Multimedia*, early access, Nov. 11, 2021, doi: 10.1109/TMM.2021.3125130.

[26] S. Kosugi, T. Yamasaki, and K. Aizawa, "Object-aware instance labeling for weakly supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6064–6072.

[27] D. Zhang, W. Zeng, J. Yao, and J. Han, "Weakly supervised object detection using proposal- and semantic-level relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3349–3363, Jun. 2022.

[28] Q. Jia, S. Wei, T. Ruan, Y. Zhao, and Y. Zhao, "GradingNet: Towards providing reliable supervisions for weakly supervised object detection by grading the box candidates," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, pp. 1682–1690.

[29] Z. Wu, J. Wen, Y. Xu, J. Yang, X. Li, and D. Zhang, "Enhanced spatial feature learning for weakly supervised object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 8, 2022, doi: 10.1109/TNNLS.2022.3178180.

[30] G. Yan et al., "C-MIDN: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9834–9843.

[31] Y. Xu, C. Zhou, X. Yu, B. Xiao, and Y. Yang, "Pyramidal multiple instance detection network with mask guided self-correction for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3029–3040, 2021.

[32] W. Gao, F. Wan, J. Yue, S. Xu, and Q. Ye, "Discrepant multiple instance learning for weakly supervised object detection," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108233.

[33] Z. Ren et al., "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10598–10607.

[34] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao, “Cyclic guidance for weakly supervised joint detection and segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 697–707.

[35] X. Li, M. Kan, S. Shan, and X. Chen, “Weakly supervised object detection with segmentation collaboration,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9735–9744.

[36] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye, “Min-entropy latent model for weakly supervised object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1297–1306.

[37] D. Zhang, J. Han, L. Zhao, and T. Zhao, “From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5549–5560, Dec. 2020.

[38] T. Cao, L. Du, X. Zhang, S. Chen, Y. Zhang, and Y.-F. Wang, “CaT: Weakly supervised object detection with category transfer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3070–3079.

[39] B. Dong, Z. Huang, Y. Guo, Q. Wang, Z. Niu, and W. Zuo, “Boosting weakly supervised object detection via learning bounding box adjusters,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2876–2885.

[40] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, “C-MIL: Continuation multiple instance learning for weakly supervised object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2199–2208.

[41] L. Sui, C.-L. Zhang, and J. Wu, “Salvage of supervision in weakly supervised object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14227–14236.

[42] G. Ghiasi, T.-Y. Lin, and Q. V. Le, “NAS-FPN: Learning scalable feature pyramid architecture for object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.

[43] M. Tan, R. Pang, and Q. V. Le, “EfficientDet: Scalable and efficient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[44] C. Lin, S. Wang, D. Xu, Y. Lu, and W. Zhang, “Object instance mining for weakly supervised object detection,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11482–11489.

[45] Y. Shen, R. Ji, K. Yang, C. Deng, and C. Wang, “Category-aware spatial constraint for weakly supervised detection,” *IEEE Trans. Image Process.*, vol. 29, pp. 843–858, 2020.

[46] A. Shrivastava, A. Gupta, and R. Girshick, “Training region-based object detectors with online hard example mining,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

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

[48] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[49] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2018, pp. 3–19.

[50] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, “An empirical study of spatial attention mechanisms in deep networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6688–6697.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[52] J. R. R. Uijlings, E. A. Van De Sande Koen, G. Theo, and W. M. S. Arnold, “Selective search for object recognition,” *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[53] J. Pont-Tuset, P. Arbeláez, J. T. Barron, F. Marques, and J. Malik, “Multiscale combinatorial grouping for image segmentation and object proposal generation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, Jan. 2017.

[54] Y. Wei et al., “TS2C: Tight box mining with surrounding segmentation context for weakly supervised object detection,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 434–450.

[55] A. Arun, C. V. Jawahar, and M. P. Kumar, “Dissimilarity coefficient based weakly supervised object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9432–9441.

[56] K. Yang, P. Zhang, P. Qiao, Z. Wang, D. Li, and Y. Dou, “Objectness consistent representation for weakly supervised object detection,” in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1688–1696.

[57] K. Yang, D. Li, and Y. Dou, “Towards precise end-to-end weakly supervised object detection network,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8372–8381.

[58] Z. Chen, Z. Fu, R. Jiang, Y. Chen, and X.-S. Hua, “SLV: Spatial likelihood voting for weakly supervised object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 12995–13004.

[59] Y. Yin, J. Deng, W. Zhou, and H. Li, “Instance mining with class feature banks for weakly supervised object detection,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3190–3198.

[60] M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis, “C-WSL: Count-guided weakly supervised localization,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 152–168.

Zhihao Wu received the B.S. degree from the Department of Computer Science, Harbin Engineering University, Harbin, China, in 2019. He is currently pursuing the Ph.D. degree with the Harbin Institute of Technology, Shenzhen, China. His research interests include computer vision and machine learning, especially weakly supervised learning.

Chengliang Liu (Graduate Student Member, IEEE) received the B.S. degree in computer science from Jilin University, Changchun, China, in 2018, and the M.S. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2020. He is currently pursuing the doctoral degree with the Harbin Institute of Technology, Shenzhen, China. His research interests include machine learning and computer vision, especially multiview representation learning.

Jie Wen (Member, IEEE) received the Ph.D. degree in computer science and technology from the Harbin Institute of Technology, Shenzhen. His research interests include biometrics, pattern recognition, and machine learning.

Yong Xu (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2005. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen. His current interests include pattern recognition, biometrics, machine learning, and video analysis.

Jian Yang (Member, IEEE) received the Ph.D. degree from the Nanjing University of Science and Technology in 2002. In 2003, he was a Postdoctoral Researcher at the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow at the Biometrics Centre, The Hong Kong Polytechnic University. From 2006 to 2007, he was a Postdoctoral Fellow at the Department of Computer Science, New Jersey Institute of Technology. He is currently a Chang-Jiang Professor with the School of Computer Science, NUST. His research interests include pattern recognition, computer vision, and machine learning. He is a fellow of IAPR.

Xuelong Li (Fellow, IEEE) is currently a Full Professor with the School of Artificial Intelligence, OPtics and ElectroNics, Northwestern Polytechnical University, Xian, China.