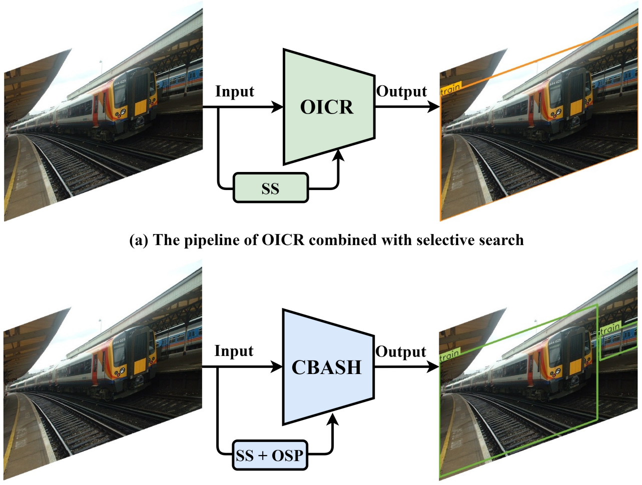
CBASH: Combined Backbone and Advanced Selection Heads with Object Semantic Proposals for Weakly Supervised Object Detection

Ruiyang Xia, Guoquan Li, *Member, IEEE,* Zhengwen Huang, *Member, IEEE,* Hongying Meng, *Senior Member, IEEE,* and Yu Pang

***Abstract*—Most recent object detection methods have achieved growing performance on public datasets. However, enormous efforts are needed for these methods due to the extensive annotations of ground-truth boxes. Weakly Supervised Object Detection (WSOD) methods hence have been proposed to solve this problem as only image-level annotations are required and then output bounding boxes related to the objects. In order to further elevate the weakly supervised detection methods on the extraction of reasonable features, the training of potential positive proposals, and the generation of proposals before training, we propose a new Combined Backbone and Advanced Selection Heads (CBASH) method with the proposals generated from the object semantic information. Specifically, Combined Backbone will make the unobvious object features more noticeable, Advanced Selection Heads promote more potential positive proposals to get training, and the generated object semantic proposals elevate the quality and quantity of positive proposals. The proposed method is evaluated on the challenging PASCAL VOC 2007 and 2012 benchmark datasets. Experimental results show that our proposed method can achieve improved performance on both VOC 2007 and VOC 2012 datasets and outperforms the existing state-of-the-art methods.**

***Index Terms*—Weakly Supervised Object Detection, Image-level annotations, Combined Backbone, Advanced Selection Heads, Object semantic proposals.**



(a) The pipeline of OICR combined with selective search

(b) The pipeline of CBASH combined with selective search and object semantic proposal

Fig. 1. Comparison of the detection result between OICR and our scheme. “SS” and “OSP” indicate selective search and object semantic proposal method, respectively. Note that the green and orange predicted boxes indicate good and weak detection results, respectively.

I. Introduction

With the amazing performance achieved by Convolutional Neural Network (CNN) model on multiple computer vision tasks such as image classification [1]–[3], object detection [4]–[10], and image segmentation [11]–[13], more advanced neural networks with various strengths are proposed to improve further. However, in object detection, fully supervised methods with massive anchor boxes [4]–[6] or anchor points [7]–[9] are in the majority, which compel researchers to put much attention on precisely annotating the coordinates of ground-truth boxes for each object before training.

Therefore, to circumvent the demand for ground-truth box annotations, Weakly Supervised Object Detection (WSOD) has been proposed in recent years as only image-level annotations are required. For WSOD models, all proposals are trained on image-level annotations and the reasonable ones will then be selected after training, which means fewer resources are needed, but achieve or even outperform the performance with massive ground-truth boxes offering. Currently, many WSOD models employ Multiple Instance Learning (MIL) [14] and end-to-end model structures to build the connections between image-level labels and proposals in an image. For example, as an effective WSOD model, Online Instance Classifier Refinement (OICR) [15] firstly train all proposals with limited image-level information and then introduce pseudo image-level ground-truth information to further focus on partial proposals.

Although the weakly supervised detection performance gets improvement greatly, three problems are still existed in these methods. Firstly, for the proposals, the quality and quantity

This work was supported by the National Key Research and Development Program of China under Grant 2019YFC1511300; the National Natural Science Foundation of China under Grant 61971079; the Basic Research and Frontier Exploration Project of Chongqing under Grant cstc2019jcyj-msxmX0666 and the Innovative Group Project of the National Natural Science Foundation of Chongqing under Grant cstc2020jcyj-cxttX0002. (Corresponding author: Guoquan Li.)

R. Xia and G. Li are with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China (e-mail: S190101135@stu.cqupt.edu.cn and ligq@cqupt.edu.cn).

Z. Huang and H. Meng are with Department of Electronic and Electrical Engineering, Brunel University London, London, UB8-3PH, United Kingdom (e-mail:Zhengwen.Huang@brunel.ac.uk and hongying.meng@brunel.ac.uk).

Y. Pang is with Key Laboratory of Photoelectric Information Sensing and Transmission Technology, Chongqing, 400065, China (e-mail: pangyu@cqupt.edu.cn).

of positive ones are poor and few, and it impedes model convergence during training. Secondly, for the features, the extracted object features from the backbone exhibit high responses only in the local discriminative regions, which results in the predicted proposals located in the evident regions generating higher confidence than the others representing a complete object. Thirdly, for the detection branches, it is unreasonable that each class only has one pseudo ground-truth box as there might be more objects with the same class with different locations in an image.

Taking into account the above issues, we propose a creative method including three parts. Firstly, to elevate the quality and enlarge the number of proposals, we propose a novel way to generate proposals based on object semantic features instead of pixel-level information such as Selective Search [16] or Edge Box [17] to further help our network better convergence. Secondly, to prevent the network from overfitting to local object discriminative regions, Combined Backbone (CB) is designed to blend the features from the masked and non-masked branches. The non-masked branch is to find local discriminative parts and localize objects roughly and the masked branch focuses on finding u nobvious features. The response of unobvious features can be improved after combining these two independent branches. Thirdly, to resolve only one pseudo ground-truth box generated in each category, Zhang et al. [10] uses multiple pseudo ground-truth boxes generated from the WSOD model in each class. We propose an efficient network head called Advanced Selection Heads (ASH) to find more pseudo ground-truth boxes with high confidence, so that more potential positive proposals will be trained. However, different from [10], we do not use several instance information about the non-missed objects to correct the pseudo ground-truth information as we focus on WSOD instead of missing bounding-boxes object detection. Fig. 1 quantitively illustrates the difference between OICR and our proposed methods.

Our main contributions are as follows:

- We propose a recurrent masking method to generate proposals based on object semantic information, which can enhance the quantity and quality of positive proposals to help our network better convergence during training.
- We propose a new network backbone structure called Combined Backbone to make the unobvious features more noticeable, which can avoid the detection results being dominated by the local evident object features.
- We propose an efficient network head called Advanced Selection Heads to generate more pseudo ground-truth boxes with high confidence for each category, which can make more relevant proposals be trained.
- We evaluate the effectiveness of our method on PASCAL VOC 2007 and 2012 datasets, respectively. Extensive experimental results on different evaluation metrics demonstrate that our model can obtain a large improvement compared to other state-of-the-art methods.

II. Related Works

A. Weakly Supervised Object Detection

Multiple Instance Learning (MIL) is firstly proposed in [14] and then applied to WSOD, where each image is defined as a bag and each proposal in an image is called an instance. In WSOD, only label of a bag is given instead of the multiple instances. Therefore, instances share the same label and the models output the instances which are most relevant to the bag label.

Besides, the efficiency of adapting CNN for the task of computer vision indicates its strong ability to extract image local features [1]. Therefore, various WSOD models are proposed based on CNN with the MIL idea. Bilen and Vedaldi [18] proposed an end-to-end learned architecture called Weakly Supervised Deep Detection Network (WSDDN). They generate proposals by adopting Selective Search and Edge Box, respectively. In the detection part, they design a network with two branches which is motivated from [19] for the sake of building the interaction between the classes and proposals by multiplication. However, the problem is that all proposals are trained equivalently whether they contain an object or not and the rate of background proposals always accounts for the majority. Therefore, training all proposals disturbs the model to distinguish the foreground and background and hard to focus on real foregrounds. Moreover, the existence of local discriminative parts to objects leads to the final predicted proposals always locate in the evident parts of an object instead of the complete region.

To handle these problems, further various enhanced models are proposed. Kantorov et al. [20] proposed ContextLocNet to make use of spatial information. ContextLocNet establishes the connections between proposals and their contexts so that the network will output more precise predicted proposals. Diba et al. [21] propounded Weakly Supervised Cascaded Convolutional Networks (WSCCN). WSCCN concentrates on promoting the quality of proposals by adopting Class Activation Map (CAM) [22] and weakly supervised segmentation in a cascaded structure. Tang et al. [15] designed a novel WSOD network named OICR. The improvement compared to WSDDN is that relevant potential positive proposals can be trained independently as the generation of pseudo ground-truth boxes. Besides the coarse selection in WSDDN, OICR introduces a series of independent parallel fully connected layers at the end of the network and choose the top score proposal in each category as the pseudo ground-truth box. Each proposal can be trained again if the overlapped area with these pseudo ground-truth boxes is larger than a threshold. Nevertheless, it is unreasonable that each category only has a single pseudo ground-truth box. Thus, to choose more reliable pseudo ground-truth boxes, Tang et al. [23] further proposed Proposal Cluster Learning (PCL), which uses k-Means [24] algorithm to split proposal scores into different clusters and the top score ranking proposals are chosen if the generated cluster contains the highest score proposal. Then, these proposals will be trained if Intersection-over-Union (IoU) between the cluster centers and proposals is larger than a threshold. However, this method is complex, not only for the clustering algorithm,

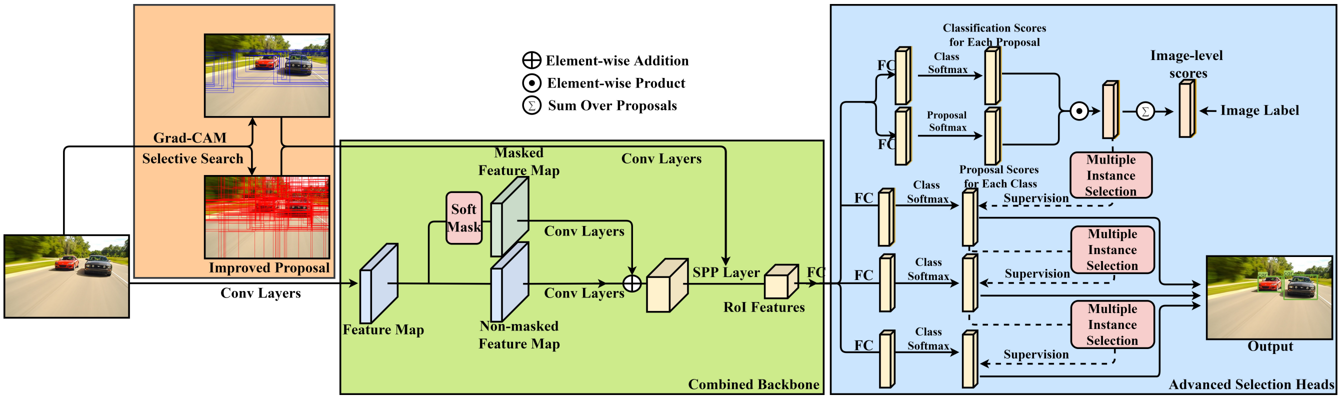


Fig. 2. The pipeline of our model is composed of three core modules. The first module is to combine pixel-level and object semantic proposals and we define the module as Improved Proposal. The second module is to decrease the influence of evident features by elevating the responses of unobvious object features. We call this module Combined Backbone (CB). The last module called Advanced Selection Heads (ASH) is to generate more pseudo ground-truth boxes to train more potential reliable proposals.

but also for the loss function with the marginal improvement compared to the performance of OICR.

Moreover, there are some other works that incorporate novel structure [25], weakly supervised image segmentation [26], and other optimizations such as sidestepping local optima [18], [27]–[29], improving the quality of the proposal [30], [31] and incorporating a regression module [32].

B. Weakly Supervised Object Localization

Similar to WSOD, due to the labor-intensive and time-consuming manual annotation work, Weakly Supervised Object Localization (WSOL) has also become a popular research area. It achieves locating objects with only image-level annotations. According to [33], the differences between WSOL and WSOD can be summarized as three points. To the goals, WSOL aims to output the coordinates of predicted boxes but WSOD still needs to output the corresponding categories. To the dataset, each input image of WSOD has multiple objects with the same or different classes. For WSOL, although each input image has only a single object, the number of learning categories far outweighs WSOD. To the evaluation metrics, Top-1/Top-5 localization accuracy (Top-1/Top-5 Loc) and localization accuracy with known ground-truth class ($GTLoc$) are used to evaluate the performance of WSOL models. For WSOD, mean Average Precision (mAP) with different IoU thresholds and Correct Localization ($CorLoc$) are used to evaluate the detection performance.

Since the effectiveness of visualized algorithms in neural network [22], this method was adapted to WSOL as the similarities existed in both vision tasks. Wei et al. [34] adopted an iterative way to locate the object by erasing part of the object regions in the input image. Zhang et al. [35] converted to erase the feature map from the CNN. Choe et al. [36] presented an attention-based dropout layer (ADL), which randomly erases evident regions in the feature map without iterative processes. Mai et al. [37] propounded Erasing Integrated Learning (EIL) which trains the erasing stream and original stream with two loss functions respectively and adopt them when the network needs to locate the object.

III. Proposed Method

In this section, we firstly introduce the process to generate object semantic proposals. Then, we describe the whole network structure (CB and ASH) in detail. The pipeline of our model is shown in Fig. 2.

A. Object Semantic Proposal

For Improved Proposal module, although the selective search method can generate mountains of proposals, the number of negative proposals far outweigh the positive ones. Besides, it is common to find many positive proposals only include small part of an object. To deal with these problems, we raise the quantity and quality of positive proposals according to object semantic features.

Object semantic proposals are produced by analyzing the object semantic features generated from the model that has already owned a certain knowledge representation. Gradient-weighted CAM (Grad-CAM) is a visualized algorithm that aims to show the response of an object with a specific class. The algorithm is more practical than CAM [22] and their identical effect can be proved mathematically [38]. To be specific, Grad-CAM can make the model output the object response by using the gradients from backpropagation for each category and adopting the global average pooling function to compute the average weights for each feature map. Then, these average weights are multiplied with the corresponding value of the feature map before ReLU function. The whole process can be summarized as follow:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k} \tag{1}$$

$$L_c = \mathrm{ReLU}\left(\sum_k \alpha_k^c A_k \right) \tag{2}$$

Where y_c is the output of class c. A_{ij}^k indicates the value in a specific position of the k-th feature map and Z denotes the area of feature map. α_k^c is the average response for the k-th feature map and L_c is the class-specific activation map.

Citation information: DOI10.1109/TCSVT.2022.3168547,

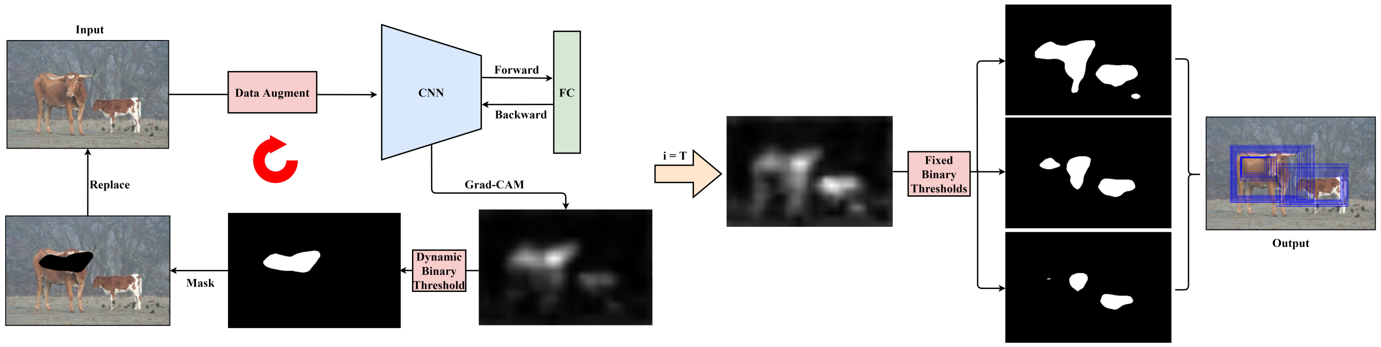


Fig. 3. The overview of the pipeline to generate object semantic proposals. The object response can be generated by using Grad-CAM. The dynamic threshold function aims to generate a mask to erase evident features. The former image is replaced by the masked one and the process will begin again. When the number of recurrent processes attains the stage threshold, multiple responses are combined to show the comprehensive result. Then the fixed threshold function filters out noises and produces the object regions. Raising the threshold will make the regions becomes smaller. Proposals are produced according to these object regions, respectively.

The differences compared with the work in [31] which also uses Grad-CAM as the base to generate object proposals can be concluded as a more generalized training-free model with fewer manual factors. To be specific, we only use VGG16 [2] pretrained on ImageNet [39] as the backbone to generate object semantic features and it can directly be utilized without training the 1000 different categories, which means our method is more generalized and can be used in the dataset with more categories. Then, compared with the ten thresholds in [31], we can also get reasonable proposals by using fewer manual thresholds based on our proposed recurrent masking method and the generation of object semantic proposals.

As the high responses are always dominated by evident features, it is important to mask these regions before generating feasible proposals. We hence adopt a recurrent process to mask these discriminative parts. The general process is illustrated in Fig. 3.

The whole process can be divided into three stages named as masking stage, combining stage, and generating stage, respectively. Firstly, the masking stage aims to erase the local discriminative parts of the object. When the original image goes through the neural network, the output will have scores related to the corresponding classes. We choose the classes within the top-5 scores as the source of responses before computing gradients via backpropagation and generated the network response before going to the dynamic binary threshold function. Then, the evident features of the object will be masked in the original image if these generated responses in a specific location are higher than the threshold from the dynamic function. After masking these evident feature regions, the masked image will replace the former one and go through the next loop.

The recurrent times we set is 2 as the first loop is to find the most discriminative parts of objects and the next is to find the secondary regions. More loops may work but we found the effect is unobvious and time-consuming as the above steps have already detected most of the regions related to objects. To the dynamic binary threshold function, the threshold value follows Eq. (3), where $i = 1, 2, \ldots, T$. T_i denotes the threshold of the i-th loop.

$$T_i = 100 + 20 * (i - 1) \tag{3}$$

According to massive empirical observations from the images, we found that a low threshold will lead to an oversized mask that includes multiple objects. 100 is a suitable benchmark to generate a reasonable mask in the first stage. As the former loop erases the discriminative parts, which means the rest of the object regions are small and the network is more sensitive to the noise when processing the masked image in the network. Therefore, a higher threshold in the next loop can avoid the network being interfered with by noises such as the background.

The combining stage is to fuse these network responses generated in each loop on average. Specifically, the comprehensive response should go through the fixed thresholds to generate the multiple sets of binary object regions as the base of proposals. As the small objects may exist in a large object region, raising the threshold can generate small object regions included in a large one so that the proposals related to these small objects can be produced. Nevertheless, it is a tradeoff that a small interval leads to the unobvious difference for each region, and a large interval results in the generated regions only representing part of the same object. Specifically, in Fig. 3, some small regions will be seen as noisy regions, which means these regions only represent part of objects and we should not consider their influence.

The generating stage is to produce proposals based on the generated object regions. However, we do not acquire proposals only from the edge of object regions. The differences between adjacent object regions are the main factor for us to produce object semantic proposals. The coordinates of the generated proposal can be computed as follows:

$$P_{right,i} = R_{right} + i * \frac{\Delta_{right}}{N} * f\left(\Delta_{right}, \Delta_{left}\right) \tag{4}$$

$$P_{left,i} = R_{left} + i * \frac{\Delta_{left}}{N} * f\left(\Delta_{left}, \Delta_{right}\right) \tag{5}$$

$$P_{top,i} = R_{top} + i * \frac{\Delta_{top}}{N} * f\left(\Delta_{top}, \Delta_{bottom}\right) \tag{6}$$

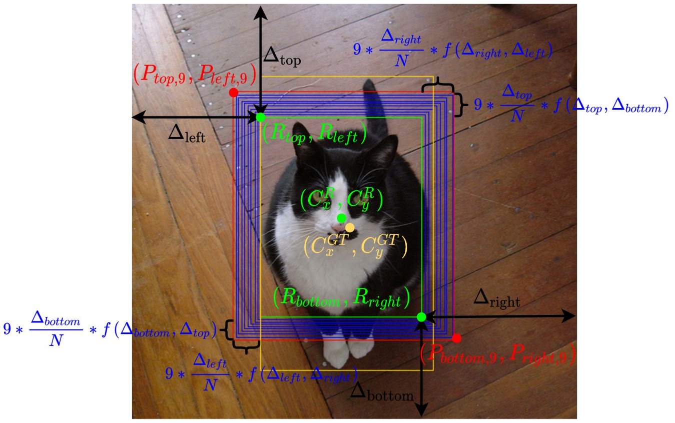

Δ_{top}
$9 * \frac{\Delta_{right}}{N} * f(\Delta_{right}, \Delta_{left})$
$(P_{top,9}, P_{left,9})$
$9 * \frac{\Delta_{top}}{N} * f(\Delta_{top}, \Delta_{bottom})$
Δ_{left}
(R_{top}, R_{left})
(C_x^R, C_y^R)
(C_x^{GT}, C_y^{GT})
(R_{bottom}, R_{right})
Δ_{right}
$9 * \frac{\Delta_{bottom}}{N} * f(\Delta_{bottom}, \Delta_{top})$
$(P_{bottom,9}, P_{right,9})$
$9 * \frac{\Delta_{left}}{N} * f(\Delta_{left}, \Delta_{right})$
Δ_{bottom}

Fig. 4. The visualized meaning of all variables in Eq. (4)-(8) and the center of the object region and the related ground-truth box. For clearness, we here only show the computation of the tenth object semantic proposal.

TABLE I
THE PROPORTION OF THE AREA OF RELATED OBJECT REGION IS LARGER THAN THE AREA OF GROUND TRUTH BOX

	IoU > 0	IoU > 0.5
$S_R < S_{GT}$	77.4%	62.7%

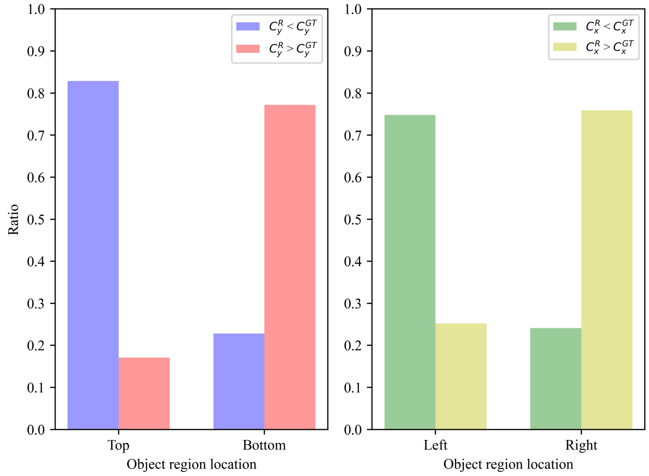


Fig. 5. The statistical results of the relative position on PASCAL VOC 2007 test set between the center of object region and the center of related ground-truth box.

$$P_{bottom,i} = R_{bottom} + i * \frac{\Delta_{bottom}}{N} * f\left(\Delta_{bottom}, \Delta_{top}\right) \quad (7)$$

$$f\left(\Delta_a, \Delta_b\right) = \begin{cases} uniform(0.7, 0.9), & \text{if } \frac{\Delta_a}{\Delta_b} > t \\ uniform(0.9, 1.0), & \text{otherwise} \end{cases} \quad (8)$$

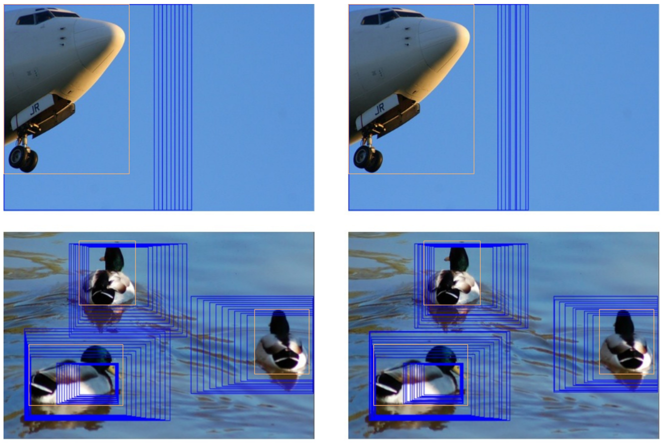
Where $P_{right,i}$, $P_{left,i}$, $P_{top,i}$ and $P_{bottom,i}$ indicate the right, left, top and bottom coordinate of i-th proposal respectively. R_{right}, R_{left}, R_{top} and R_{bottom} represent the coordinates of a generated object region before. N is a constant which represents the number of proposals for each region and t is the threshold which determines the restriction factor of the interval distance. Δ_{right}, Δ_{left}, Δ_{top} and Δ_{bottom} are the differences computed on the corresponding location. We visually show the meaning of all variables in Eq. (4)-(8) and also illustrate the center of the object region and the ground-truth box respectively in Fig. 4.

From the statistical results in Table I on PASCAL VOC 2007 test subset, S_R and S_{GT} indicate the area of generated object region and the area of its related ground-truth box, respectively. It can be noticed that although the object region has a high correlation (i.e., IoU(R, GT)>0.5) with the ground-truth box, there is still 62.7% confidence to determine the area of generated object region is smaller than the area of the related ground-truth box. We hence need to enlarge the object regions to get reasonable object semantic proposals.

The effective object region has a high correlation with a ground-truth box or insides the related ground-truth box. If an effective object region locates at the top or left of a ground-truth box, the rest of the regions must locate at the bottom or right. In WSOD, we do not have any instance-level information. However, as validated by the statistical results in Fig. 5, no matter where the effective object region is in the image, there is at least 70% probability that its center is farther from the image center than the center of the related ground-truth box, which implies that generating object semantic proposal towards the center of the image will be more reasonable. Therefore, given an effective object region located in the top or left of the image center, paying more attention on enlarging the bottom or right of the object generated region is more likely to reduce the distance between the object region center and ground-truth box center.

Furthermore, to decrease the influence of the object regions that are larger than the ground-truth boxes and bring irreverlant information, the random restriction function f in Eq. (8) is hence incorporated into Eq. (4)-(7) to shrink the growth of the coordinates, so that the irrelevant information can be reduced. Besides, objects located at the edge of the image lead to imbalanced intervals in computing the coordinates of object semantic proposals. For example, if an object is located in the top-left corner of an image, the result of Δ_{bottom} and Δ_{right} will be much larger than Δ_{top} and Δ_{left}. Background information from the bottom-right area will hence be included. The random restriction function also considers this situation and brings relatively strong restriction to the growth of the bottom-right interval. Fig. 6 illustrates the comparisons for object semantic proposals by adopting the random restriction function. It can be noticed that the object semantic proposals are relatively tight to the ground-truth boxes and less irrelevant information is included compared to the generated proposals without using the random function. However, as most object regions are smaller than their related ground-truth boxes, stronger restriction in Eq. (8) will lead to more proposals with smaller areas.

As we stated above, the adjacent sets of object regions have connections that decide the result of Δ. The former object regions will become the boundary of the latter ones. But for the first object regions, the size of the original image acts as their farthest boundary. Moreover, there is a potential problem that the shrinking regions sometimes represent the same object rather than other small objects in Fig. 3, which means these



(a) Semantic proposals without random function **(b) Semantic proposals with random function**

Fig. 6. Comparisons between the object semantic proposals without and with the random restriction function. The orange and blue boxes indicate the ground-truth boxes and object semantic proposals, respectively. The generated proposals with the random restriction function are relatively tighter to the ground-truth boxes, which means the background information within these proposals can be reduced.

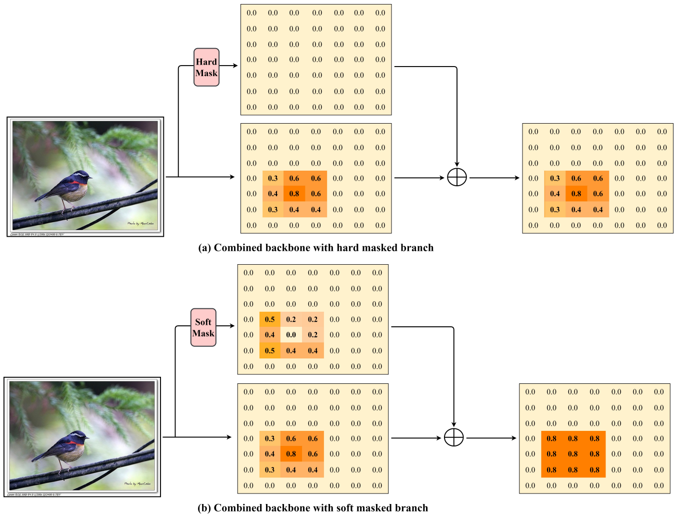
regions should not be considered during the generation of object semantic proposals. To decrease the influence, we compute the IoU between the former object region and the latter ones to decide whether to generate proposals based on these smaller regions.

In addition, the image pyramid is also introduced as each local feature is computed within a local receptive field by CNN, which means the responses are changed by the scale of an object. Thus, we set $s = [1, 0.75, 0.5]$ and suggest that it is suitable for the network to find small, middle, and large size objects, respectively. The initial threshold in the fixed binary function also equals 100, to generate more sets of object regions, each raising threshold behind the function we set will be higher than 30 compared with the last one, i.e., the fixed binary thresholds $T_{fix} = [100, 130, 160]$. Besides, each object region will generate a fixed number of proposals. According to Eq. (4)-(7), when the Δ is computed, the interval distance is decided by N. Here, we only choose the top 10 proposals (i.e., $i \in [0, 9]$) from each object region and set N to 50 and 20 for the first and the rest sets of the object regions respectively. Lastly, the threshold t we set is 8 in Eq. (8).

B. Combined Backbone

Given an image, proposals based on pixel-level and object semantic features are generated and fed into the Spatial Pyramid Pooling (SPP) layer [40]. Specifically, after getting the extracted features from CB and the coordinates of proposals, following [18] and [15], we map the proposals to the feature map in order to get the Region of Interests (RoIs). As each RoI has its own size, SPP layer is then adopted to get the RoIs with a fixed size by replacing the max pooling layer located in the last of CB. Then, the parallel fully connected layers in ASH will get them and output a series of the proposal feature vectors.

Toward further decreasing the overfitting of the local evident features. A combined network structure is designed to elevate the responses of unobvious feature regions. Specifically, after extracting the low-level features, the network will be divided into two independent branches. The first branch is to find the discriminative parts of an object to roughly localize the position and the second one is to mask these discriminative parts but keep the unobvious features.



(a) Combined backbone with hard masked branch

(b) Combined backbone with soft masked branch

Fig. 7. Difference between hard and soft masked method. Here we assume that α equals 0. After going through the soft masked method, the feature values around the object are the same, which means all features surrounding the object have the same level of importance. However, the hard masked method does not work as the average feature value is decreased by a too large background area.

The way to delimit the discriminative and unobvious features is to extract the mean and maximum from the feature maps before going through these two separate branches. The discriminative features will be masked if the value is higher than a threshold and the response of unobvious features will be reinforced. Moreover, there are two methods to implement value determination. One is the hard masked method and the other is the soft masked one which can be shown in Eq. (9) and Eq. (10), respectively.

$$f^{h}_{i,j} = \begin{cases} 0, & \text{if } f_{i,j} > mean + \alpha * (Max - mean). \\ f_{i,j}, & \text{otherwise.} \end{cases} \tag{9}$$

$$f^{s}_{i,j} = \begin{cases} Max - f_{i,j}, & \text{if } f_{i,j} > mean + \alpha * (Max - mean) \\ f_{i,j}, & \text{otherwise} \end{cases} \tag{10}$$

Where $f_{i,j}$ denotes the feature value in a specific position. $f^{h}_{i,j}$ and $f^{s}_{i,j}$ indicate the result by adopting the hard and soft masked method respectively. α is a hyperparameter that controls the determination of threshold.

In comparison with the hard masked method, there are two advantages to the soft one. Firstly, it can avoid a situation where some features whose value is slightly higher than the threshold can not get improvement as the hard masked function will set zero to them in the masked branch. Secondly, it can work well in the situation that small objects locate in a large background as too much background information cause the reduction of the global average feature value so that the masked branch can only focus on features with a much low

Algorithm 1 Multiple Instance Selection

```
Input: Outputs $S$; threshold $t$; threshold $K$; image label vector $\mathbf{y} = [y_1, \dots, y_C]$; proposals $P$; index vector $\mathbf{c} = [1, \dots, C]$.
Output: Pseudo ground-truth boxes $\hat{G} = [\hat{g}_1, \dots, \hat{g}_C]$
Randomize the sequence of $\mathbf{c}$ and assign to the original class vector.
for $i = 1$ to $|\mathbf{c}|$ do
    if $y_{c[i]} == 1$ then
        Initilize $k$.
        Sort $S$ from high to low.
        Select the top 10% proposals from $P$ as $P'$.
        for $j = 1$ to $|P'|$ do
            if $k < K$ then
                if $k == 0$ or $\mathrm{IoU}\left(P'[j], \hat{g}_{c[i]}\right) < t$ then
                    $\hat{g}_{c[i]}$.append($P'[j]$).
                    $k = k + 1$.
                    Clear the score of $P'[j]$.
            else
                Break.
return $\hat{G}$
```

value for hard masked method. In Fig. 7, we generally show a simple example to clearly illustrate the difference between the function of the hard and soft masked method.

C. Advanced Selection Heads

The procedure of ASH can be summarized into two steps. All proposals generated in the Improved Proposal module are firstly trained in the top head to find potential foreground proposals. The ground-truth image-level labels are also fed as supervisions. Then, only potential positive proposals will be trained again in the rest of the refinement heads, which is similar to the correction operation in [5]. The pseudo ground-truth box labels generated in the former head are fed in the next head as new supervisions. Different from OICR [15] or PCL [23] which produce purely one or multiple pseudo ground-truth boxes with complex clustering algorithms for each category, a simple and efficient algorithm called Multiple Instance Selection (MIS) is proposed to generate multiple pseudo ground-truth boxes.

MIS algorithm aims to generate multiple pseudo ground-truth boxes with high confidence in each class and send them to the latter refinement head so that more relevant feasible proposals can be trained. The details of MIS are shown in Algorithm 1. Based on the outputs from a network head branch, proposals have different scores in each class. We sort all of the scores from high to low and focus on the top 10% score proposals P' as the source of pseudo ground-truth boxes. A threshold K will be set to restrain the number of pseudo ground-truth boxes in each category. After that, we adopt a similar way as non-maximum suppression (NMS) with the same threshold in each class to determine whether a proposal can act as the pseudo ground-truth box. More specifically, for i-th index, if the IoUs between a high score proposal and the pseudo ground-truth boxes generated before are lower than the threshold t, this proposal will be appended to the list of pseudo ground-truth boxes $\hat{g}_{c[i]}$. Note that the scores of these selected proposals are cleared as these proposals might be trained in other classes. Moreover, the sequence of class vector $\mathbf{c}$ is randomized before generating the pseudo ground-truth boxes $\hat{G}$ to decrease the influence that the class with prior order can select more proposals.

In ASH, each class will generate multiple pseudo ground-truth boxes. The proposals whose the largest IoU computed among these pseudo ground-truth boxes higher and lower than 0.5 will be treated as foregrounds and backgrounds, respectively. To train our network more efficiently, balancing the positive and negative proposals is needed as the number of backgrounds is far outweigh the foreground. Therefore, we multiply the standard softmax loss with the weight γ^k in the k-th head such that the loss function for foreground and background samples can be balanced by adjusting the value of γ^k. To avoid the model overfitting by this parameter, all fine branches have the same adjustment of γ. Our experiments show that this weight can significantly elevate the model performance. Moreover, we follow [15] that multiplies the w_p which equals the score of the closest pseudo ground-truth box for proposal p because these boxes in their corresponding class as the supervisions are very noisy and w_p is small at the beginning of training. This variable can prevent our model from being disturbed by these noisy pseudo boxes. Therefore, for the k-th refinement head, the refinement loss function is shown in Eq. (11)-(13).

$$FG_{c,p}^k = w_p^k y_{c,p}^k \log s_{c,p}^k \tag{11}$$

$$BG_p^k = w_p^k y_{(C+1)p}^k \log s_{(C+1)p}^k \tag{12}$$

$$L_R^k = -\frac{1}{|P|} \sum_{p=1}^{|P|} \left(\gamma^k \sum_{c=1}^{C} FG_{c,p}^k + \left(1 - \gamma^k\right) BG_p^k \right) \tag{13}$$

Where $FG_{p,c}^k$ and BG_p^k indicate the forground with c-th class and background cost for p-th proposal in k-th head, respectively. C is the number of class and $|P|$ equals the number of proposals. $C+1$ represents the label of background. $y_{c,p}^k$ and $s_{c,p}^k$ indicate the label and output of the class c for the proposal p in k-th head.

Then, the standard binary cross entropy function which is shown in Eq. (14) from the top corse selection head is incorporated into the whole model loss function to make the network trained jointly. The whole loss function used to train our network is shown in Eq. (15).

$$L_{WSDDN} = -\sum_{c=1}^{C} y_c \log P_c + (1 - y_c) \log (1 - P_c) \tag{14}$$

$$L = L_{WSDDN} + \sum_{k=1}^{K} L_R^k \tag{15}$$

Where P_c and y_c indicate the output of the top head and ground-truth image-level label for the c-th class. C and K represent the number of class and refinement heads, respectively.

TABLE II
PER-CLASS DETECTION RESULTS USING VGG16 ON PASCAL VOC 2007.

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	mAP_{50}
ContextLocNet [20]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
Li [28]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	21.8	57.1	22.1	34.4	50.5	61.8	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
OICR [15]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
Self-taught [30]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	54.3	3.0	53.6	24.7	43.6	48.4	65.8	6.6	18.8	51.9	43.6	53.6	62.4	41.7
WSCCN [21]	49.5	60.6	38.6	29.2	16.2	**70.8**	56.9	42.5	10.9	44.1	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	60.8	54.5	42.8
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
TS2C [27]	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 [25]	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
Shen [26]	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
Wan [41]	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
Yi [32]	62.1	67.9	**51.6**	22.3	18.4	69.3	68.0	47.9	23.1	54.9	42.2	49.0	51.3	67.3	13.0	24.0	46.6	53.1	61.8	58.9	47.6
C-MIL [29]	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
PG-PS (extra training) [31]	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**
Ours	**65.5**	**74.4**	47.5	**36.3**	22.8	67.7	68.3	25.9	**26.9**	63.9	37.0	30.2	52.5	**70.4**	3.1	27.6	**58.9**	51.6	62.5	61.5	47.7

TABLE III
PER-CLASS CORRECT LOCALIZATION RESULTS USING VGG16 ON PASCAL VOC 2007

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	$CorLoc$
Li [28]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
ContextLocNet [20]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	47.1	59.5	67.0	83.5	35.3	39.9	67.0	49.7	63.5	65.2	55.1
Self-taught [30]	72.7	55.3	53.0	27.8	35.2	68.6	81.9	60.7	11.6	71.6	29.7	54.3	64.3	88.2	22.2	53.7	72.2	52.6	68.9	75.5	56.1
WSCCN [21]	83.9	72.8	64.5	44.1	40.1	65.7	82.5	58.9	33.7	72.5	25.6	53.7	67.4	77.4	26.8	49.1	68.1	27.9	64.5	55.7	56.7
OICR [15]	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	81.4	60.6
TS2C [27]	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
Wan [41]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.4
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
WSRPN [25]	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
Shen [26]	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
Yi [32]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**-**	-	65.0
C-MIL [29]	82.1	75.7	73.0	44.2	43.5	76.7	83.6	75.9	40.7	76.7	44.5	68.8	77.9	88.0	41.8	54.6	68.0	58.9	74.9	74.2	66.2
PG-PS (extra training) [31]	**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**
Ours	85.3	**87.9**	63.5	54.1	42.5	83.3	86.0	35.7	**47.0**	85.8	52.6	56.4	68.2	85.6	9.4	58.4	**86.6**	**72.2**	68.7	76.9	65.8

IV. EXPERIMENTS

In this section, we firstly introduce the datasets and evaluation metrics utilized in experiments. Training strategies and implementation details are described in the second part. Then, we compare the performance of our method with other state-of-the-art ones. Extensive experiments are conducted in the next part to discuss the influence of different parameter settings and modules. Finally, some qualitative results are shown to further illustrate the effectiveness of our method.

A. Datasets and Evaluation Metrics

We evaluate our model on the challenging PASCAL VOC 2007 and 2012 datasets following standard PASCAL VOC protocol [42]. These two benchmarks are split into train, validation, and test sets, which contain 9,963 and 22,531 images for 20 object classes respectively. Here we combine train and validation sets as trainval set for both benchmarks and choose them (5,011 images for 2007 and 11,540 for 2012) respectively with only image-level annotations to train our network. In the testing stage, Average Precision (AP) and mAP with IoU threshold at 50% and 75% are both utilized to evaluate the model performance. Moreover, $CorLoc$ [43] is also taken as the evaluation metric to measure the localization accuracy on the trainval and test set, respectively.

B. Training Strategies and Implementation Details

During training, Adam [44] is adopted as the optimizer of our network model with the initialized learning rate equals 1×10^{-5}. The mini-batch size we set is 8 on VOC 2007 and 16 on VOC 2012 respectively. For both datasets, we train the model for 30k iterations and decrease the learning rate by 0.1 after 20k iterations. If not specified, γ^k equals 0.7 for positive proposals in the refinement loss function for each refinement branch and α equals 0.1 in the soft masked function. In algorithm 1, K and threshold t are set to 5 and 0.05 as default. The number of refinement branches is 3. With the random horizontal flipping operation, we utilize five image scales $\{480, 576, 688, 864, 1200\}$ to resize the shortest side of the training image stochastically. In the testing stage, the outputs are computed by analyzing all scales and their horizontal flips for each image comprehensively.

We use Selective Search to generate pixel-level proposals combined with our object semantic proposals as the training samples. VGG16 (without batch normalization) pretrained on the ImageNet [39] dataset is adopted as the backbone of feature extraction. Besides, the penultimate max-pooling layer and subsequent convolutional layers are changed to dilated convolutional layers to enlarge the receptive field. For the newly added layers, they are initialized by using Gaussian distributions with 0-mean and standard deviations of 0.01. Biases are initialized to 0.

TABLE IV
PER-CLASS DETECTION RESULTS USING VGG16 ON PASCAL VOC 2012.

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	mAP_{50}
ContextLocNet [20]	64.0	54.9	36.4	8.1	12.6	53.1	40.5	28.4	6.6	35.3	34.4	49.1	42.6	62.4	19.8	15.2	27.0	33.1	33.0	50.0	35.3
Li [28]	62.9	55.5	43.7	14.9	13.6	57.7	52.4	50.9	13.3	45.4	4.0	30.2	55.6	67.0	3.8	23.1	39.4	5.5	50.7	29.3	35.9
WSCCN [21]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.9
OICR [15]	67.7	61.2	41.5	**25.6**	22.2	54.6	49.7	25.4	19.9	47.0	18.1	26.0	38.9	67.7	2.0	22.6	41.1	34.3	37.9	55.3	37.9
Self-taught [30]	60.8	54.2	34.1	14.9	13.1	54.3	53.4	58.6	3.7	53.1	8.3	43.4	49.8	69.2	4.1	17.5	43.8	25.6	55.0	50.1	38.3
Shen [26]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.1
Wei [27]	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 *
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
WSRPN [25]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.8
Wan [41]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	42.4
Yi [32]	69.5	68.3	**53.1**	17.4	27.7	55.5	53.5	45.3	19.8	60.1	26.9	47.7	54.8	72.0	**24.5**	26.2	51.1	31.3	58.3	56.0	45.9
C-MIL [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.7
PG-PS (extra training) [31]	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**
Ours	**72.8**	**71.6**	47.1	22.6	**31.2**	59.6	57.8	34.8	**24.8**	**62.3**	19.1	54.7	**66.2**	**75.6**	3.7	**27.9**	**57.3**	44.2	48.8	59.1	47.0*

http://host.robots.ox.ac.uk:8080/anonymous/I2CJJP.html

TABLE V
PER-CLASS CORRECT LOCALIZATION RESULTS USING VGG16 ON PASCAL VOC 2012

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	TV	$CorLoc$
Li [28]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	29.1
ContextLocNet [20]	78.3	70.8	52.5	34.7	36.6	80.0	58.7	38.6	27.7	71.2	32.3	48.7	76.2	77.4	16.0	48.4	69.9	47.5	66.9	62.9	54.8
Self-taught [30]	82.4	68.1	54.5	38.9	35.9	84.7	73.1	64.8	17.1	78.3	22.5	57.0	70.8	86.6	18.7	49.7	80.7	45.3	70.1	77.3	58.8
OICR [15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62.1
PCL [23]	77.2	83.0	62.1	**55.0**	49.3	83.0	75.8	37.7	43.2	81.6	46.8	42.9	73.3	90.3	21.4	56.7	84.4	55.0	62.9	82.5	63.2
Shen [26]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.5
Wei [27]	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 [25]	85.5	60.8	62.5	36.6	53.8	82.1	80.1	48.2	14.9	**87.7**	**68.5**	60.7	**85.7**	89.2	**62.9**	**62.1**	**87.1**	54.0	45.1	70.6	64.9
C-MIL [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.4
Yi [32]	84.6	79.9	**73.7**	42.8	53.1	83.7	69.2	72.0	47.8	84.8	51.5	64.7	78.5	90.3	43.8	55.1	81.9	46.5	73.6	79.8	67.9
PG-PS (extra training) [31]	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**
Ours	**89.0**	**87.0**	67.1	48.0	**55.8**	86.5	77.5	57.6	**54.5**	86.8	44.4	63.8	82.4	**92.4**	13.9	58.1	85.8	**67.3**	68.2	79.5	68.3

Our experiments are implemented in Pytorch [45] deep learning framework, Python, and C++. All of the experiments run on NVIDIA GTX TitanV GPU and Intel Xeon Silver 4110 CPU (2.10 GHz).

C. Comparsion with State-of-the-Art

We calculate the AP_{50} and mAP_{50} for each and all categories on the PASCAL VOC 2007 test set, respectively. Then, $CorLoc$ is computed on the trainval set to compare our method with other state-of-the-art ones. Experimental results are shown in Table II and Table III,

It can be noticed that our proposed method can achieve the comparable results on PASCAL VOC 2007. Specifically, our model gets the best performance in the class of aeoplane, bike, boat, chair, motor, and sheep for AP_{50}, and bike, chair, sheep, and sofa for $CorLoc$ on VOC 2007. However, our method can only get 3.1% AP and 9.4% $CorLoc$ to the category of person, which is weaker than previous methods. We conjecture the reasons for the generated inferior performance can be attributed to two parts. Firstly, we only use PASCAL VOC 2007 trainval set to train our model, which means there are fewer data compared to PASCAL VOC 2012 trainval set. It leads to the CBASH can not be trained very well as the CB needs data to know the unobvious object features and elevate their responses and the ASH needs data to generate reliable multiple pseudo ground-truth boxes in each class and train the potential positive proposals better. Moreover, if the number of samples is small, the number of object semantic proposals will be small and then offer benefits to a limited extent. Secondly, the overfitting existed in pretrained VGG16 will decrease the quality of these proposals. As shown in Fig. 8, after we mask the evident features of the person, the network still generates a high response around the evident features instead of the unobvious features such as hands or feet.

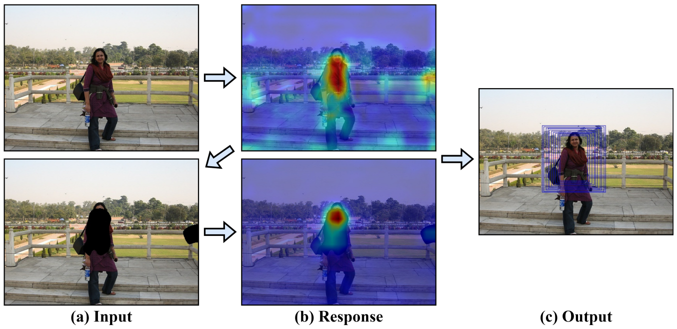

(a) Input
(b) Response
(c) Output

Fig. 8. Object semantic proposals about person generated via VGG16 model pretrained on ImageNet. After masking the discriminative region such as the face, the pretrained model still generates a high response in the region.

More data can help our model own stronger knowledge representation and more precise localization. We then report the corresponding evaluated results to analyze the model performance on PASCAL VOC 2012. Results are listed in Table IV and Table V, respectively. It can be observed that

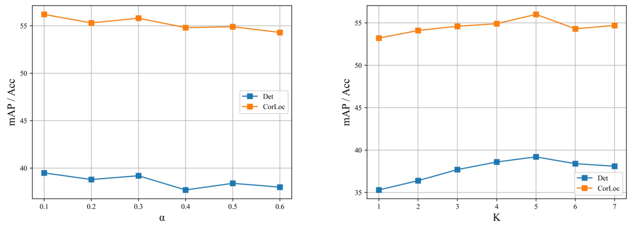


Fig. 9. Effect of α (left) and K (right) on The PASCAL VOC 2007 Dataset.

we win in 9 categories of AP_{50} and attain the significantly shrink the gap of mAP_{50} compared to [31] which uses an extra training classifier. We may suggest the performance can be attributed to more positive proposals with high qualities being produced, the network learning more unobvious features, and more potential positive proposals can be found and get training. Also, we get more comparable comprehensive performance on $CorLoc$ and win the categories on aeroplane, bike, bottle, chair, motor and sofa, respectively. To be specific, we significantly outperform the second-best one over 6.7% in the sheep samples (54.5% vs. 47.8%).

D. Discussion and Ablation Studies

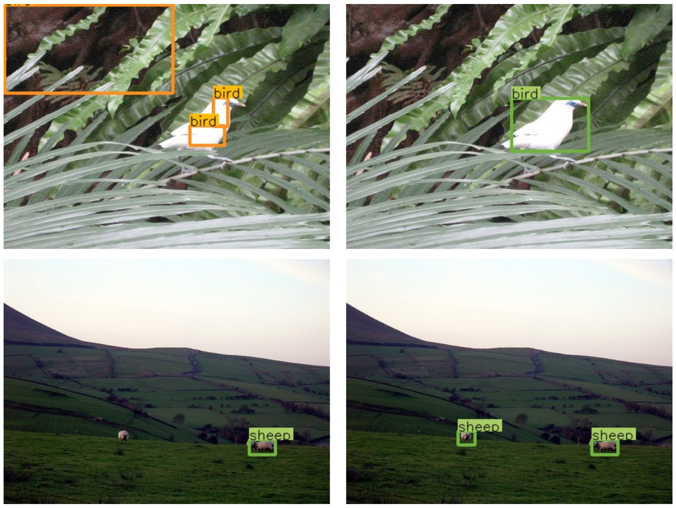
In this part, we analyze the influence of various model hyperparameters which have already been elucidated in the last section, i.e., γ^k, α, and K without the influence of object semantic proposals and multiple image scales during training. Then we discuss the performance of different modules in our model, i.e., object semantic proposals, CB, and ASH through the ablation studies.

We firstly discuss the influence of γ^k, this hyperparameter is to adjust the unbalance of positive and negative samples as there are amounts of proposals generated from IP module in each image and the negative proposals are in the majority. The results are shown in Table VI. We can find that when the value of γ^k is 0.9, mAP_{50} and mAP_{75} do not get the highest values, which we suggest that the restriction of negative samples is too strong. After changing γ^k from 0.9 to 0.7, they increase steadily and achieve the maximum when γ^k equals 0.7. Then, continue decreasing the γ^k leads to the degradation of performance as the unbalance enlarged between the foreground and background.

TABLE VI
EFFECT OF γ^k ON THE PASCAL VOC 2007 DATASET.

γ^k	mAP_{50}	mAP_{75}
0.90	37.3	12.1
0.85	37.5	12.2
0.80	38.2	12.4
0.75	38.7	12.6
0.70	**39.5**	**13.2**
0.65	38.9	12.8

Then we discuss the influence of α, the hyperparameter is to control the detemination of threshold in the feature map. The higher α leads to larger masked regions. Results are



(a) Hard masked method **(b) Soft masked method**

Fig. 10. Comparisons of qualitative detection results between (a) the hard masked method and (b) the soft masked method.

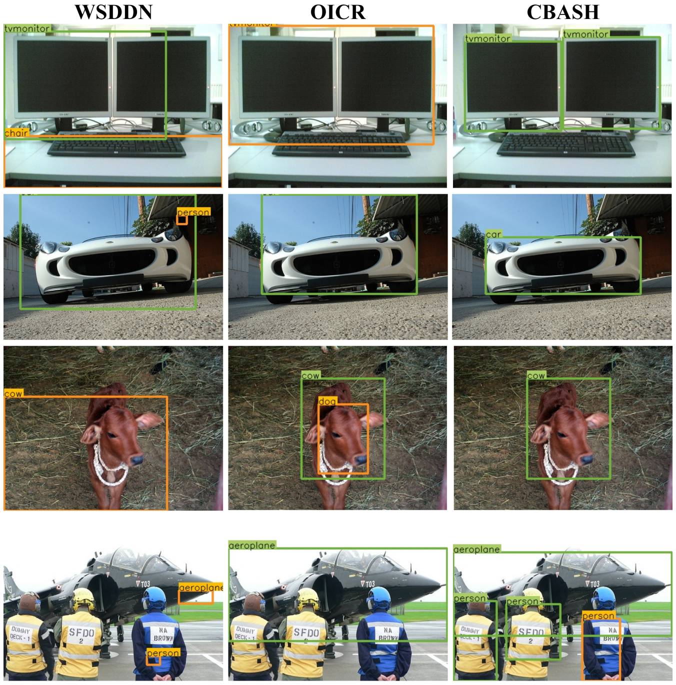


Fig. 11. Comparisons of qualitative results among WSDDN (left), OICR (middle) and CBASH (right) on PASCAL VOC 2007.

illustrated on the left of Fig. 9. From this line graph, we observe that their maximum appeared when α equals 0.1, and the overall tendency for the curve of $CorLoc$ and mAP_{50} are decreasing with the increasing number of α, which means there is a suitable combination between the masked and non-masked branch in CB and the noisy features gradually impact the model when increasing α.

Next, we discuss the influence of K in our network. It is responsible for assigning the maximum of pseudo ground-truth boxes in each category. Related experiments are also shown on the right of Fig. 9. According to this graph, it can be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCSVT.2022.3168547,

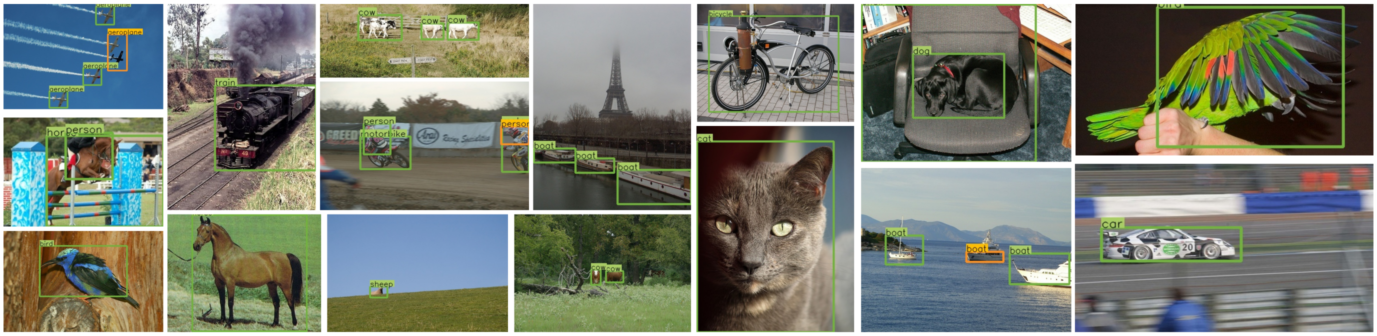


Fig. 12. Some visualized detection results for class aeroplane, train, cow, boat, bicycle, dog, bird, person, horse, cat and car.

apparently noticed that both curves have a similar tendency, and setting K to 5 can obtain the best performance. Note that more pseudo ground-truth boxes can not bring the gain of detection performance, which means ensuring the quality of pseudo ground-truth boxes is important during training.

TABLE VII
COMPARISONS OF DETECTION RESULTS AMONG DIFFERENT COMBINATIONS IN BACKBONE ON THE PASCAL VOC 2007

Combination	mAP_{50}	mAP_{75}	$CorLoc$
CB with only non-masked branch	36.7	12.1	52.5
CB with hard masked method	38.8	12.7	54.3
CB with soft masked method	**39.5**	**13.2**	**56.2**

To demonstrate the effectiveness of the proposed soft masked method, the detection results by employing only the non-masked branch, the hard masked method and the softed masked method in CB are shown in Table VIII together. It can be noticed that either the hard or soft masked method can boost the detection performance compared to the only non-masked branch in CB, which means the enhanced responses of unobvious features can detect the objects more precisely. Furthermore, as the soft masked method can also notice the features whose values are slightly larger than the threshold and the small objects with a large background, the detection performance of the soft masked method is hence better than the hard masked one. Moreover, to further demonstrate the strength of the soft masked method which detects small objects in a large background, we collect the images from the VOC 2007 test set whose ground-truth boxes of objects are smaller than one-third of the image sizes. Some visualized results are shown in Fig. 10 and the detection results are listed in Table VIII. It can be seen from this Table that the soft masked method can respectively outperform the detection performance over 2.8% and 4.2% on mAP_{50} and $CorLoc$. According to Fig. 10, the soft masked method can not only produce more precise predicted boxes, but also find small objects more effectively than the hard masked one, which presents the strength of the soft masked method is better than the hard masked method in detecting small objects.

TABLE VIII
COMPARISONS OF THE SMALL OBJECTS DETECTION RESULTS BETWEEN THE HARD AND SOFT MASKED METHOD ON THE COLLECTED PASCAL VOC 2007

Combination	mAP_{50}	$CorLoc$
CB with hard masked method	21.8	41.3
CB with soft masked method	**24.6**	**45.5**

In the ablation studies, we respectively discuss the improved extent of each module to the baseline model. The results are listed in Table IX. It can be observed that all components can elevate the performance of baseline especially for our ASH which boosts the performance at 4.2% for mAP_{50}. After incorporating object semantic proposals, the performance of CBASH can further be elevated and attain the best performance on mAP_{50} (47.7%) and $CorLoc$ (65.8%), which demonstrate the effectiveness of the object semantic proposals.

TABLE IX
ABLATION STUDIES ON THE PASCAL VOC 2007.

Object Semantic Proposals	Combined Backbone	Advanced Selection Heads	mAP_{50}	$CorLoc$
			41.2	60.6
✓			42.3	61.2
	✓		43.1	61.8
		✓	45.4	63.1
	✓	✓	46.1	63.8
✓	✓	✓	**47.7**	**65.8**

E. Qualitative Results

Toward further illustrate the effectiveness of CBASH, we compare our model with WSDDN [18] and OICR [15] qualitatively. All methods utilize the same basic network with identical training strategies and we do not incorporate object semantic proposals during training to ensure fairness. The results are shown in Fig. 11. Note that we determine the quality of predicted boxes according to whether the IoU between the predicted box and ground-truth box is larger than 0.5. From these visualized results, we can observe that our model can localize the object more precisely without producing extra false positive predicted boxes. For example, in the third row of Fig. 11, a little cow is similar to the class of dog to some extent. Our model can detect and recognize it precisely but the others can not localize well or even generate wrong predicted boxes. Moreover, when multiple objects with different classes appeared tightly in an image, our model can still detect them.

Although our model outperforms the state-of-the-art methods and is robust to the viewpoints, scales, and occlusions for the objects with different classes, it is also challenging to detect them completely for some special categories such as person, which will be our future work to research from the generation of more advanced proposals and the design of more generalized network during training. More qualitative results are shown in Fig. 12.

V. Conclusion and Future Work

In this paper, we proposed a new network model called CBASH with object semantic proposals to further elevate WSOD performance. We improve the quantity and quality of the positive proposals, the response of unobvious features, and the generation of pseudo ground-truth boxes by using our object semantic proposals, Combined Backbone (CB), and Advanced Selection Heads (ASH), respectively. Specifically, the Improved Proposal module offers more positive proposals generated from object semantic information to help our model better convergence during training. The CB module elevates the response of unobvious features to avoid the network focusing too much on the local discriminative regions. The ASH module produces more pseudo ground-truth boxes in each category so that more potential relevant proposals will be trained again. Extensive experiments demonstrate that our method can bring a great improvement and significantly outperform other state-of-the-art methods. For the future, the adaptive threshold value and the improved network response can be incorporated to further elevate the quality of proposals. Moreover, the advanced coordinate regression loss function can be introduced to correct the coordinates of proposals and elevate the detection performance.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[3] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2021.

[4] Xiaoyu Chen, Hongliang Li, Qingbo Wu, King Ngi Ngan, and Linfeng Xu. High-quality r-cnn object detection using multi-path detection calibration network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2):715–727, 2020.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[6] Jung Uk Kim, Jungsu Kwon, Hak Gu Kim, and Yong Man Ro. Bbc net: Bounding-box critic network for occlusion-robust object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1037–1050, 2020.

[7] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[8] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019.

[10] Yongqiang Zhang, Mingli Ding, Yancheng Bai, Mengmeng Xu, and Bernard Ghanem. Beyond weakly supervised: Pseudo ground truths mining for missing bounding-boxes object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):983–997, 2019.

[11] Yiren Zhou, Thanh-Toan Do, Haitian Zheng, Ngai-Man Cheung, and Lu Fang. Computation and memory efficient image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):46–61, 2016.

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[13] Tao Lei, Peng Liu, Xiaohong Jia, Xuande Zhang, Hongying Meng, and Asoke K Nandi. Automatic fuzzy clustering framework for image segmentation. *IEEE Transactions on Fuzzy Systems*, 28(9):2078–2092, 2019.

[14] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[15] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2843–2851, 2017.

[16] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[17] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.

[18] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[19] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[20] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *European Conference on Computer Vision*, pages 350–365. Springer, 2016.

[21] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 914–922, 2017.

[22] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[23] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):176–191, 2018.

[24] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[25] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 352–368, 2018.

[26] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 697–707, 2019.

[27] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2409–2416, 2014.

[28] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016.

[29] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly super-

vised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019.

[30] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1377–1385, 2017.

[31] Gong Cheng, Junyu Yang, Decheng Gao, Lei Guo, and Junwei Han. High-quality proposals for weakly supervised object detection. *IEEE Transactions on Image Processing*, 29:5794–5804, 2020.

[32] Sheng Yi, Huimin Ma, Xi Li, and Yu Wang. Wsodpb: Weakly supervised object detection with pcsnet and box regression module. *Neurocomputing*, 418:232–240, 2020.

[33] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[34] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

[35] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018.

[36] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019.

[37] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8766–8775, 2020.

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[41] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1306, 2018.

[42] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[43] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision*, 100(3):275–293, 2012.

[44] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

Ruiyang Xia received the B.S. degree in Electronic and Information Engineering from Chongqing University of Posts and Telecommunications at Chongqing, China, in 2019. He is currently pursuing his M.S. degree in Communication and Information Engineering at Chongqing University of Posts and Telecommunications, China. His research interests include image processing, image recognition, object detection, and machine learning.

Guoquan Li received the M.S. and Ph.D. degrees in circuits and systems from Chongqing University, Chongqing, China, in 2006 and 2012, respectively. From 2009 to 2010, he was a Visiting Ph.D. Student with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA. He is currently a Professor with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications (CQUPT), Chongqing. His current research interests include image processing, machine learning and MIMO wireless networks.

Zhengwen Huang currently is a member of Brunel Institute of Power Systems / senior research fellow of SERG Brunel. He received the Ph.D. degree from the Department of Electronic and Computer Engineering at Brunel University London, UK. He received the MSc degree from King's College London and his BSc from University of Science and Technology, China. His research is focused on image processing and data engineering.

Hongying Meng (M'10-SM'17) received his Ph.D.degree in Communication and Electronic Systems from Xi'an Jiaotong University, Xi'an, China. He is an associate editor for IEEE Transactions on Circuits and Systems for Videos Technology (TCSVT) and IEEE Transactions on Cognitive and Developmental Systems (TCDS). He has authored over 170 publications including IEEE TIP, TCYB, TFS, TAC, TCSVT, TBE, TCDS, ICASSP and CVPR. He is currently a Reader at the Department of Electronic and Electrical Engineering, Brunel University London, U.K. His research interests include digital signal processing, affective computing, machine learning, human computer interaction, and computer vision.

Yu Pang received the B.S. degree in electrical engineering from Sichuan University, Sichuan, China, in 2000, the M.S. degree (honors) in communication and information engineering from the University of Electronic Science and Technology of China, Sichuan, in 2003, and the Ph.D. degree from the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada. He is currently a Professor with Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China. His current research interests include integrated circuit design, wireless communication, and artificial intelligence.