High-Quality Proposals for Weakly Supervised Object Detection

Gong Cheng, Junyu Yang, Decheng Gao, Lei Guo, and Junwei Han, *Senior Member, IEEE*

***Abstract*— Despite significant efforts made so far for Weakly Supervised Object Detection (WSOD), proposal generation and proposal selection are still two major challenges. In this paper, we focus on addressing the two challenges by generating and selecting high-quality proposals. To be specific, for proposal generation, we combine selective search and a Gradient-weighted Class Activation Mapping (Grad-CAM) based technique to generate more proposals having higher Intersection-Over-Union (IOU) with ground truth boxes than those obtained by greedy search approaches, which can better envelop the entire objects. As regards proposal selection, for each object class, we choose as many confident positive proposals as possible and meanwhile only select class-specific hard negatives to focus training on more discriminative negative proposals by up-weighting their losses, which can make training more effective. The proposed proposal generation and proposal selection approaches are generic and thus can be broadly applied to many WSOD methods. In this work, we unify them into the framework of Online Instance Classifier Refinement (OICR). Experimental results on the PASCAL VOC 2007 and 2012 datasets and MS COCO dataset demonstrate that our method significantly improves the baseline method OICR by large margins (13.4% mAP and 11.6% CorLoc gains on the VOC 2007 dataset, 15.0% mAP and 8.9% CorLoc gains on the VOC 2012 dataset, and 6.4% mAP and 5.0% CorLoc gains on the COCO dataset) and achieves the state-of-the-art results compared with existing methods.**

***Index Terms*— Weakly supervised object detection (WSOD), proposal generation, proposal selection, convolutional neural networks (CNNs).**

I. Introduction

WEAKLY Supervised Object Detection (WSOD) is the task of learning object detectors with only image-level labels that indicate the presence or absence of an object class. More recently, WSOD has attracted considerable attention because of its broad applications [1]–[12]. Compared with fully supervised object detection paradigm [13]–[20] which requires object-level annotations (i.e., object bounding boxes) at training stages, WSOD is more promising but also more challenging to obtain satisfactory performance.

Manuscript received August 15, 2019; revised March 12, 2020; accepted April 8, 2020. Date of publication April 16, 2020; date of current version April 27, 2020. This work was supported in part by the National Science Foundation of China under Grant 61772425, Grant 61773315, Grant 61790552, and Grant U1801265, in part by the Key R&D Program of Guangdong Province under Grant 2019B010110001, in part by the Fundamental Research Funds for the Central Universities under Grant 3102019AX09, and in part by the Research Funds for Interdisciplinary subject, NWPU. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Soma Biswas. *(Corresponding author: Junwei Han.)*

The authors are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

Digital Object Identifier 10.1109/TIP.2020.2987161

(a) Proposals of selective search (b) Proposals of our method

Fig. 1. Some top-scoring proposals (green boxes) measured in terms of IOU between the proposals and the ground truth boxes (yellow boxes). The proposals in (a), obtained by selective search method [21], can not well fit the ground truth boxes. The proposals in (b), obtained by our method, have higher quality than that obtained by selective search [21].

As a widely-used paradigm for weakly supervised object detection, Multiple Instance Learning (MIL) based methods have been extensively explored in recent years [2], [22]–[28]. In brief, MIL treats each training image as a bag and the image regions generated by object proposal methods as instances. Then, it formulates the task of weakly supervised object detection task as a learning problem that alternates between two steps. (i) Training object detectors (instance classifiers) based on a fixed set of training samples. (ii) Updating the training set by using the learned object detectors. Although most of existing MIL based WSOD methods have achieved promising results, there still exist two major challenges.

First, how to generate high-quality proposals? Owing to the lack of object bounding boxes annotations, MIL methods generally train object detectors with the instances obtained by proposal generation methods such as selective search [21]. Figure 1 (a) shows some top-scoring instances measured in terms of Intersection-Over-Union (IOU) between the proposals, obtained by selective search method [21], and the ground truth bounding boxes. As shown in Figure 1 (a), the proposals can not well fit the ground truth bounding boxes. Under such situation, the learned object detectors may not well or even can not localize objects considering the performance requirement of high IOU values (IOU$>$0.5 is often used for a true positive). This motivates us to generate high-quality proposals which can better envelop the entire objects, as shown in Figure 1 (b).

Second, how to select high-quality proposals? Since there is no explicit information available about how many objects from a given object class exists in each image, many existing MIL methods, such as the representative Online Instance Classifier

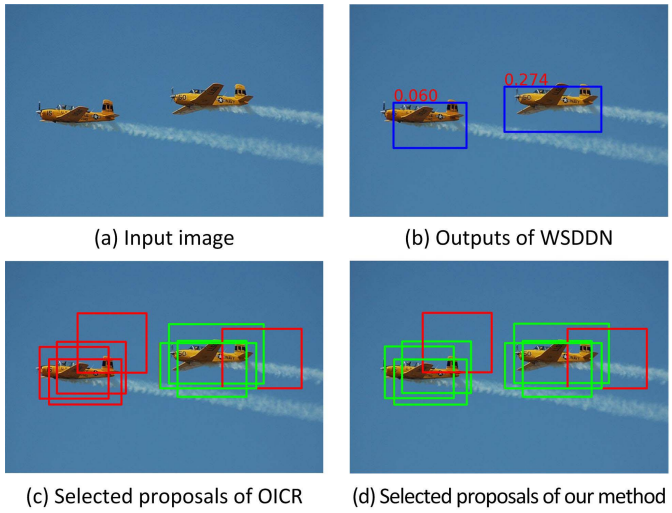


Fig. 2. Given an image (a), OICR [2] only selects the proposal with the highest score (0.274 in (b)), obtained with WSDDN [1], and its highly spatially overlapped proposals as positive proposals (green boxes) and the rest as negative proposals (red boxes). Our method aims to select as many confident positive proposals as possible and mine class-specific hard negative proposals, which can make training more effective.

Refinement (OICR) [2] which is also our baseline method, only select the proposal with the highest score along with its spatially overlapped proposals as positive instances and the rest as negative instances (see Figure 2 (c)) for object detector training. However, as shown in Figure 2, in practice, most of the real-word images usually contain multiple objects of the same class, so merely selecting green boxes as positive instances and red boxes as negative instances is problematic because this will result in omitting some very valuable positive instances and also introducing inaccurate negative instances for object detector training. This issue motivates us to adaptively select as many confident positive proposals as possible and meanwhile mine more discriminative negative proposals, as shown in Figure 2 (d).

In this paper, we focus on addressing the two challenges by generating and selecting high-quality proposals. To sum up, our main contributions are as follows. First, we present a simple proposal generation method. By combining selective search [21] and a Gradient-weighted Class Activation Mapping (Grad-CAM) [29] based technique, our method generates more proposals that have higher IOU with ground truth boxes than those obtained by greedy search approaches, which can better envelop the entire objects. Second, we propose an effective proposal selection method. Specifically, for each object class, we select as many confident positive proposals as possible and meanwhile only select class-specific hard negatives to focus training on more discriminative negative proposals by up-weighting their losses, which can make training more effective. Third, our proposed proposal generation and proposal selection approaches are generic and thus can be broadly applied to many WSOD methods. In this work, we unify them into the framework of OICR [2]. Experimental results on the PASCAL VOC 2007 and 2012 datasets and MS COCO dataset demonstrate that our method significantly improves the baseline method OICR [2] by large margins and achieves the state-of-the-art results.

II. Related Work

The challenge faced for WSOD task is how to distinguish object instances from the complex backgrounds. A branch of approaches for tackling WSOD is to formulate it as an MIL problem by treating each training image as a bag and iteratively selecting high-scoring instances from each bag when learning object detectors [27], [30]–[37].

In recent years, with the development of deep learning, especially Convolutional Neural Networks (CNNs), deep MIL networks have been actively studied due to their better performance. Bilen and Vedaldi [1] proposed a two-stream CNN weakly supervised deep detection network (WSDDN), which selects the positive samples by multiplying the scores of classification and detection. Since then, the introduction of WSDDN [1] has opened the door through deep CNN models to improve object detection performance. Built on the WSDDN, Kantorov et al. [38] introduced two kinds of context-aware guidance, including additive and contrastive models, to improve localization. Tang et al. [2] proposed an OICR approach to refine instance classifiers by propagating instance labels to spatially overlapped instances. By combining WSDDN [1] and OICR [2], Zhang et al. [39] designed a weakly-supervised to fully-supervised framework, where a weakly-supervised detector is implemented using MIL.

Besides, Wan et al. [36] proposed an effective continuation MIL (C-MIL) method by introducing a continuation optimization algorithm into MIL, with the intention of alleviating the non-convex optimization problem. Arun et al. [40] designed a novel dissimilarity coefficient based WSOD framework, which achieves the WSOD task via minimizing the difference between an annotation agnostic prediction distribution and an annotation aware conditional distribution. Yang et al. [37] designed a unified end-to-end network with a MIL branch and a bounding-box regression branch, together with an attention module, that share the same backbone. Shen et al. [41] presented a WSOD framework termed Weakly Supervised Joint Detection and Segmentation (WS-JDS) by combining the tasks of weakly supervised object detection and segmentation into a multi-task learning framework.

Our method is also related to proposal generation and refinement such as the works of objectness [42], tight box mining with surrounding segmentation context [43], region context refinement [44], soft proposal networks [45], and Self-Produced Guidance (SPG) network [46]. Besides, our work is partially related to sample selection and sample mining, such as Focal Loss [47] and especially Online Hard Example Mining (OHEM) [48]. However, different from OHEM which mines the training examples according to the loss of each example based on fully supervised learning where each image is annotated with object bounding boxes, our method selects the samples based on the proposals' scores and the image labels. Also, we propose a simple but effective up-weighting scheme to assign larger weights to hard negatives to facilitate the model training, which shares similar idea with Focal Loss to some extent.

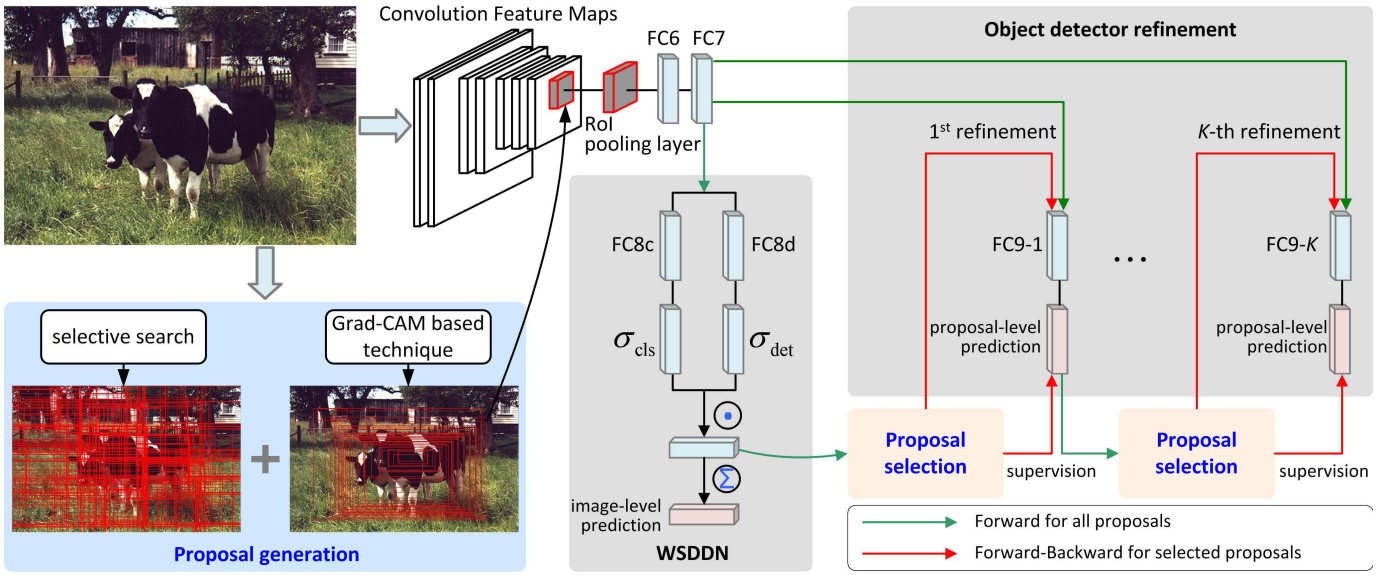


Fig. 3. The architecture of our proposed method, which is built on OICR framework [2] by introducing two important modules including proposal generation and proposal selection. Here, proposal generation aims to obtain more high-quality proposals that have higher IOU with ground truth boxes. Proposal selection aims to adaptively select more confident positive proposals and mine discriminative hard negatives. Green arrow denotes the forward pass for all input proposals and red arrow represents the selected proposals which are input to the object detector refinement networks.

Among all methods mentioned above, the OICR work [2] is most relevant to our method. However, this paper mainly focuses on addressing the two challenges of proposal generation and proposal selection via the framework of OICR [2]. Thus, our method can be easily inserted into most of existing MIL-based WSOD methods.

III. Methodology

As we stated above, this paper mainly focuses on generating and selecting high-quality proposals to further boost the performance of WSOD. To this end, we implement our method based on the state-of-the-art OICR [2] framework by introducing two important modules including proposal generation and proposal selection, as shown in Figure 3. To be specific, proposal generation aims to obtain more high-quality proposals that have higher IOU with ground truth boxes and thus can better fit the entire objects. Proposal selection is an iterative operation that aims to adaptively select more confident positive proposals and meanwhile mine discriminative hard negatives for refining object detectors. Besides, in order to make training more effective, we propose to up-weigh the losses of class-specific hard negatives to focus training on more discriminative proposals.

A. Proposal Generation

As we have analyzed before and illustrated in Figure 1, most of the proposals generated by selective search [21] can not well fit the ground truth bounding boxes of objects, which results in that the learned detectors may not well localize objects. So we propose to generate more high-quality proposals by combining selective search [21] and a Grad-CAM [29] based technique. Since selective search is a mature technique for object proposal generation, we here mainly focus on describing how to generate high-quality object proposals by means of Grad-CAM [29]. Specifically, we use a coarse-to-fine, two-stage strategy to generate object proposals.

The first stage uses VGG16 model [49] to train a set of coarse classifiers with image-level labels for the task of multi-label image classification with the sigmoid cross-entropy loss function as follows:

$$S = -\sum_{i=1}^{C}(y_i \log P_i + (1 - y_i)\log(1 - P_i)) \tag{1}$$

where C is the total number of image classes, y_i is the label indicator for the i-th image class, and P_i is the prediction result of the i-th sigmoid classifier.

For each image containing object class c, with the coarse classifiers, we can obtain its class-specific activation map M_c by performing a weighted combination of a set of convolutional feature maps, and follow it by a ReLU:

$$M_c = \text{ReLU}(\sum_k \alpha_k^c \mathbf{A}_k) \tag{2}$$

where $\mathbf{A}_k$ is the k-th convolutional feature map. α_k^c is the 'importance' of feature map $\mathbf{A}_k$ for an object class c, which is computed by globally-average-pooling the gradient of y_c with respect to $\mathbf{A}_k$ as follows:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y_c}{\partial (\mathbf{A}_k)_{ij}} \tag{3}$$

where y_c is the score of the c-th classifier before sigmoid.

For each class-specific activation map of a given input image, we first set ten segmentation thresholds equally distributed between the maximum gray value of the activation map and the average gray value over all pixels. Then, for each

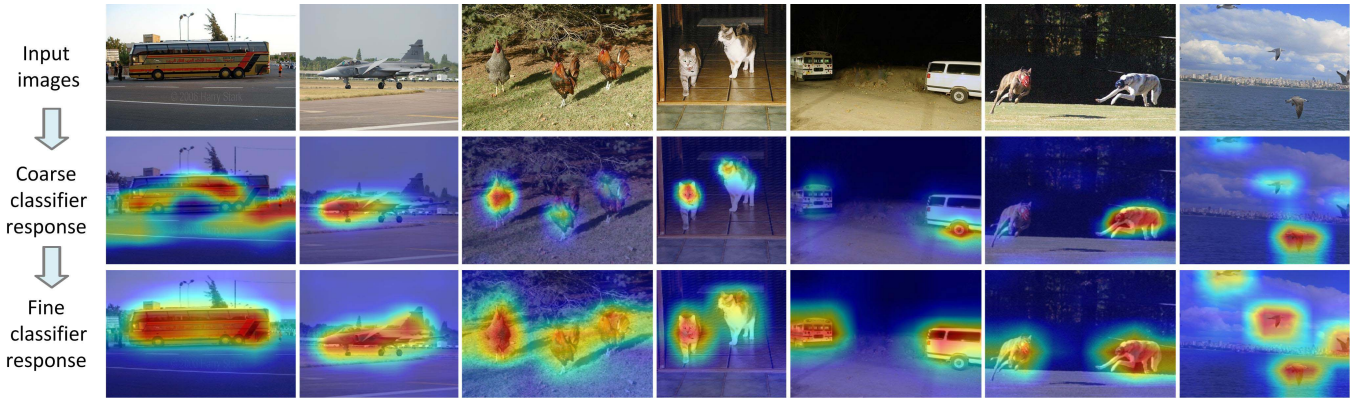


Fig. 4. Input images and their corresponding class-specific activation maps obtained by coarse classifiers and fine classifiers.

segmentation threshold, we obtain a binary image from the class-specific activation map. Finally, we use the maximum connected area method to get a set of bounding boxes, each of which tightly encloses a largest connected area, and these bounding boxes are treated as object proposals. In this way, we can obtain a large number of class-specific object proposals. However, as shown in the middle row of Figure 4, although high-response regions contain objects, they are still far from fully locating the whole objects.

To fix this issue, we need to further train a set of fine classifiers for better localizing the whole objects under weakly supervised setting. In this situation, how to select as precise pseudo-labels as possible is critical. To this end, for a given object class, we only select the proposals of the first stage, whose softmax responses are the highest or whose sigmoid scores are 1, as the input for the training of the second stage's fine classifiers with VGG16 model. This is actually an object proposal classification task. The sigmoid cross-entropy loss function, as given in Eq. 1, is used for model training. By repeating the same operation as the first stage, we can generate higher-quality object proposals (see the third row of Figure 4), which can better locate the whole objects than the stage one.

Since there are no image-level labels available for test images, we first use image-level classifier to predict the top-scoring n potential object classes for each test image and then use proposal-level classifiers to obtain the final proposals used for detection.

B. Weakly Supervised Deep Detection Network

Weakly supervised deep detection network (WSDDN) [1] is used here to obtain proposal scores for subsequent proposal selection and object detector refinement. It works as follows. Given an image $\mathbf{x}$ and its label $\mathbf{y}_i = [y_1, \ldots, y_C]$, we can obtain a list of proposals $\mathcal{R} = \{R_1, \ldots, R_{|\mathcal{R}|}\}$ by proposal generation method, where $y_c = 1$ or 0 indicates the image with or without object class c, and C is the number of object classes. Then, its object proposal features (the output of FC7) are fed into two data streams, termed classification data stream and detection data stream, to obtain two matrices of data $\mathbf{x}^c$, $\mathbf{x}^d \in \mathbb{R}^{C\times|\mathcal{R}|}$ via two fully-connected layers FC8c and FC8d, respectively. These two matrices are then passed through two softmax operators, respectively, to produce the classification scores and detection scores of each proposal as follows:

$$[\sigma_{\text{cls}}(\mathbf{x}^c)]_{ij} = \frac{e^{x^c_{ij}}}{\sum_{k=1}^{C} e^{x^c_{kj}}}, \tag{4}$$

$$[\sigma_{\text{det}}(\mathbf{x}^d)]_{ij} = \frac{e^{x^d_{ij}}}{\sum_{k=1}^{|\mathcal{R}|} e^{x^d_{ik}}}. \tag{5}$$

The final proposal scores are obtained by using the element-wise product $\mathbf{x}^{\mathcal{R}} = \sigma_{\text{cls}}(\mathbf{x}^c) \odot \sigma_{\text{det}}(\mathbf{x}^d)$, which will be used for proposal selection and object detector refinement. And then, we transform them to obtain an image-level prediction score of the c-th class, P_c, by the sum over all proposals:

$$P_c = \sum_{r=1}^{|\mathcal{R}|} x^{\mathcal{R}}_{cr}. \tag{6}$$

For WSDDN model training, stochastic gradient descent is used to optimize the following loss function over all training images by using P_c in Eq. 6 as the predicted probability belonging to the c-th image class:

$$J_1 = -\sum_{c=1}^{C} (y_c \log P_c + (1 - y_c) \log (1 - P_c)). \tag{7}$$

C. Proposal Selection

After we have obtained all proposal scores, how to select high-quality proposals adaptively is crucial for the final results. In OICR method [2], given an image with a target object class, it only selected the highest-scoring proposal of that class together with its spatially overlapped proposals as positive instances and the rest as negative instances, as shown in Figure 2 (c). However, this process is obviously problematic when faced for the images that contain multiple target objects of the same class because this will result in omitting some actual and valuable positive proposals and also introducing inaccurate negative proposals. In addition, each image often contains a large number of easy negatives and a small number of hard negatives, but the former ones contribute much less to the model updating than the latter, so using all negative

proposals to refine object detectors is very time-consuming. To address this problem, we propose a simple yet extremely effective proposal selection strategy.

More specifically, for an input image with a list of object proposals $\mathcal{R} = \{R_1, \ldots, R_{|\mathcal{R}|}\}$ and their corresponding proposal scores $\mathbf{x}^{\mathcal{R}}$, the proposed proposal selection module for each object class c works as follows.

Positive proposal selection. If $y_c = 1$, it says that the image contains at least one object of the target class c, then the proposal j_c with the highest score is selected according to Eq. 8 and is labelled to a pseudo class c, i.e., $y_{cj_c} = 1$. Inspired by the fact that the proposals having high spatial overlaps may share the same label, if the IOU between the proposals j and j_c is greater than a user-defined threshold (0.5 as OICR [2]), we label proposal j to class c ($y_{cj} = 1$). Then, we continue to pick the highest-scoring proposal (except for those positives that have been selected before) to obtain positives as described above. We repeat this process until the selected proposal with the highest score is no more bigger than τ ($\tau = 0.5$ in this paper and we will report the results by using different τ values).

$$j_c = \arg\max_r x_{cr}^{\mathcal{R}} \tag{8}$$

Negative proposal selection. Our selected negative proposals consist of two types of instances. (i) The first type is the ones that have an IOU with the highest-scoring proposals j_c in the interval [0.1, 0.5) from the images with positive label $y_c = 1$. We denote them as $\mathcal{N}_{c1}$. Thus, for any $j \in \mathcal{N}_{c1}$, we have $y_{cj} = 0$. In fact, these proposals only contain parts of objects, thus selecting them for training will further improve the location accuracy of object detectors. (ii) The second type is the ones that are definitely misclassified as object class c from the images with negative label $y_c = 0$. We call them "hard negative proposals" and denote them as $\mathcal{N}_{c2}$. Any $j \in \mathcal{N}_{c2}$ shows that its source image does not contain any object of class c, but the mined proposal j has the highest score on category c. Actually, these hard negatives $\mathcal{N}_{c2}$ are more discriminative because they are obtained under strong supervision (see Table I). On the contrary, the first type of negative proposals $\mathcal{N}_{c1}$ is obtained under weak supervision.

D. Object Detector Refinement

This module focuses on progressively refining object detectors by selecting more reliable positive proposals and more discriminative negative proposals. For each refinement, we first perform a forward pass for all input proposals (see Figure 3, green arrows) by using the previous object detectors (the first refinement is based on WSDDN [1]). Then, the proposal selection module uses the procedure described in Section III-C to select confident positives and hard negatives, which are finally input to the object detector refinement networks (see Figure 3, red arrows).

Up-weighting proposal importance. As described above, the negative proposals in $\mathcal{N}_{c1}$ are the instances that have an IOU overlap in the range of [0.1, 0.5) with the pseudo positive instances which are inevitably noisy, whereas the negative proposals in $\mathcal{N}_{c2}$ are definitely correct ones obtained from the images not containing the target objects with stronger supervision information than those in $\mathcal{N}_{c1}$. So we propose an adaptive up-weighting scheme to assign larger weights to the hard negatives in $\mathcal{N}_{c2}$ to further improve the object detector refinement as follows:

$$w_{cj}^k = \begin{cases} e^{\mathrm{IOU}_{jj_{gt}} x_{cj}^{\mathcal{R}k}}, & j \in \mathcal{N}_{c1}^k \\ e^{x_{cj}^{\mathcal{R}k}}, & j \in \mathcal{N}_{c2}^k \end{cases} \tag{9}$$

where $\mathrm{IOU}_{jj_{gt}} \in [0.1, 0.5)$ is the IOU value of proposal j and its corresponding pseudo ground truth box (i.e., proposal j_{gt}). $x_{cj}^{\mathcal{R}k}$ is the score of proposal j on object class c based on the k-th refinement, $\mathcal{N}_{c1}^k$ and $\mathcal{N}_{c2}^k$ are the negative proposals of the k-th refinement.

Then, given the refinement time K and the proposal-level supervision information, for the k-th refinement, the object detectors can be trained based on the following loss function:

$$J_2^k = -\frac{1}{|\mathcal{R}|} \sum_{j=1}^{|\mathcal{R}|} \sum_{c=1}^{C+1} w_{cj}^k w_j^k y_{cj}^k \log x_{cj}^{\mathcal{R}k} \tag{10}$$

where w_j^k is the loss weight to handle the unstable solutions caused by noisy supervision. Please see [50] for more details.

We finally unify WSDDN [1] and different object detector refinement stages into an end-to-end deep network, which is optimized with the following loss function:

$$J = J_1 + \sum_{k=1}^{K} J_2^k. \tag{11}$$

IV. Experiments

A. Datasets

We evaluate our method on three datasets as follows. (1) PASCAL VOC 2007 and VOC 2012 datasets [51]. These two datasets are the most widely used benchmarks for weakly supervised object detection. The PASCAL VOC 2007 dataset contains totally 9963 images covered by 20 object classes, where 5011 images are used for training and validation, and the remaining 4952 images are used for testing. The PASCAL VOC 2012 dataset contains a total of 22531 images, including 11540 images for training and validation, and 10991 images for testing. (2) COCO dataset [52]. This is a very challenging object detection benchmark proposed in 2014, which contains 80 object categories. The training set consists of 80k images and the validation set consists of 40k images. For PASCAL VOC datasets, we use the `trainval` split for training, and the `test` split for testing. For COCO dataset, we follow common practice to adopt the COCO `trainval35k` split (the union of training set with 80k images and a random subset with 35k images from the validation set) and report the results on the `minival` split (the remaining 5k images from validation set).

The performance is measured with average precision (AP) and the mean AP over all object classes with an IOU threshold of 0.5. Besides, we also report CorLoc, a widely used WSOD measure [50]. CorLoc measures the percentage of positive training images in which a method correctly localizes an object

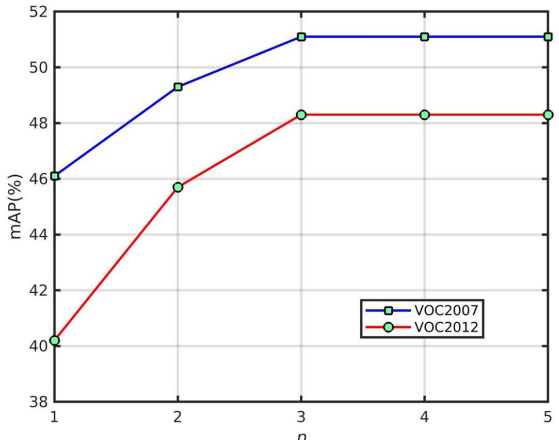


Fig. 5. Effect of n on the PASCAL VOC datasets.

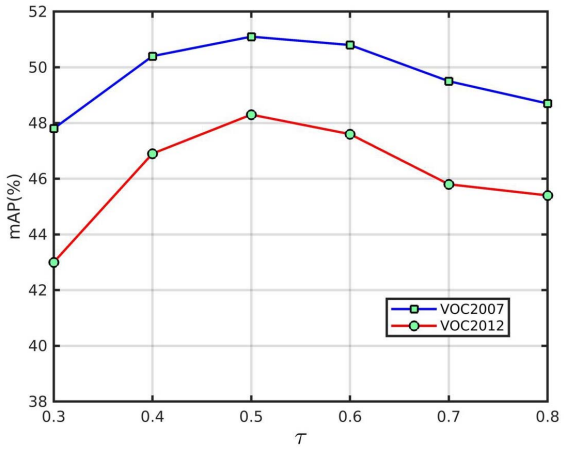


Fig. 7. Effect of τ on the PASCAL VOC datasets.

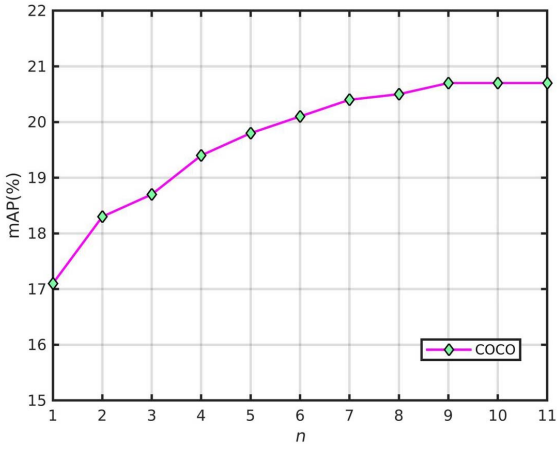


Fig. 6. Effect of n on the COCO dataset.

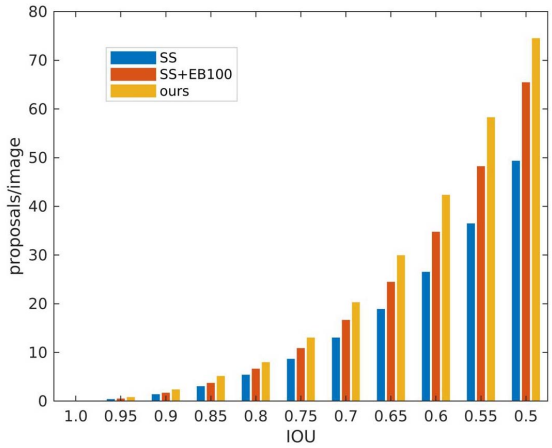


Fig. 8. IOU distribution of the proposals obtained with three different methods on the PASCAL VOC 2007 dataset.

of the target class according to the PASCAL VOC criterion (i.e., the most confident detection overlaps by at least 50% with one of the ground truth bounding boxes). CorLoc is evaluated on the union of training and validation subsets, which is different from AP that is measured on the test set.

B. Parameter Settings

Our method is built on VGG16 model [49] pre-trained on ImageNet [53] and the backbone of OICR framework [2]. We adopt the same experimental parameters as OICR [2] expect for two newly added parameters, namely n and τ. n is used to predict the top-scoring n potential object categories to obtain Grad-CAM based proposals for each test image. τ is used for the positive proposal selection. Figures 5 and 6 give the effects of n on the detection results on the PASCAL VOC datasets and the COCO dataset, respectively. As can be seen, for the PASCAL VOC datasets, $n = 3$ outperforms other choices and the results are not very sensitive to the choice of n from the set of $\{3, 4, 5\}$, while for the COCO dataset, $n = 9$ outperforms other choices and the results are also not very sensitive to the choice of n from the set of $\{9, 10, 11\}$. In our experiments, we set n to 3 and 9 for the PASCAL VOC datasets and the COCO dataset, respectively.

Figure 7 reports the influences of τ on the PASCAL VOC 2007 and 2012 datasets when τ varies from 0.3 to 0.8 with a stride of 0.1. As seen, the effects of τ have similar performance trends on both the VOC 2007 and 2012 datasets. Specifically, the accuracy first improves and then degenerates with the increase of the value of τ. The best results are achieved when $\tau = 0.5$ on both two datasets. Consequently, we empirically set τ to 0.5 in our work.

C. Ablation Studies

Our method introduces three important modules to OICR method [2]. They are proposal generation (PG), proposal selection (PS) and up-weighting scheme for hard negatives. To evaluate the contributions of the three modules, we carried out controlled experiments based on OICR framework [2] to investigate how each component affects the performance. For all the experiments, we use the same parameter settings, except for the specified changes to the component(s). The experiments are implemented on the PASCAL VOC 2007 dataset. Table I reports the results using the baseline method OICR+selective search (SS) [2] and its combination with various components.

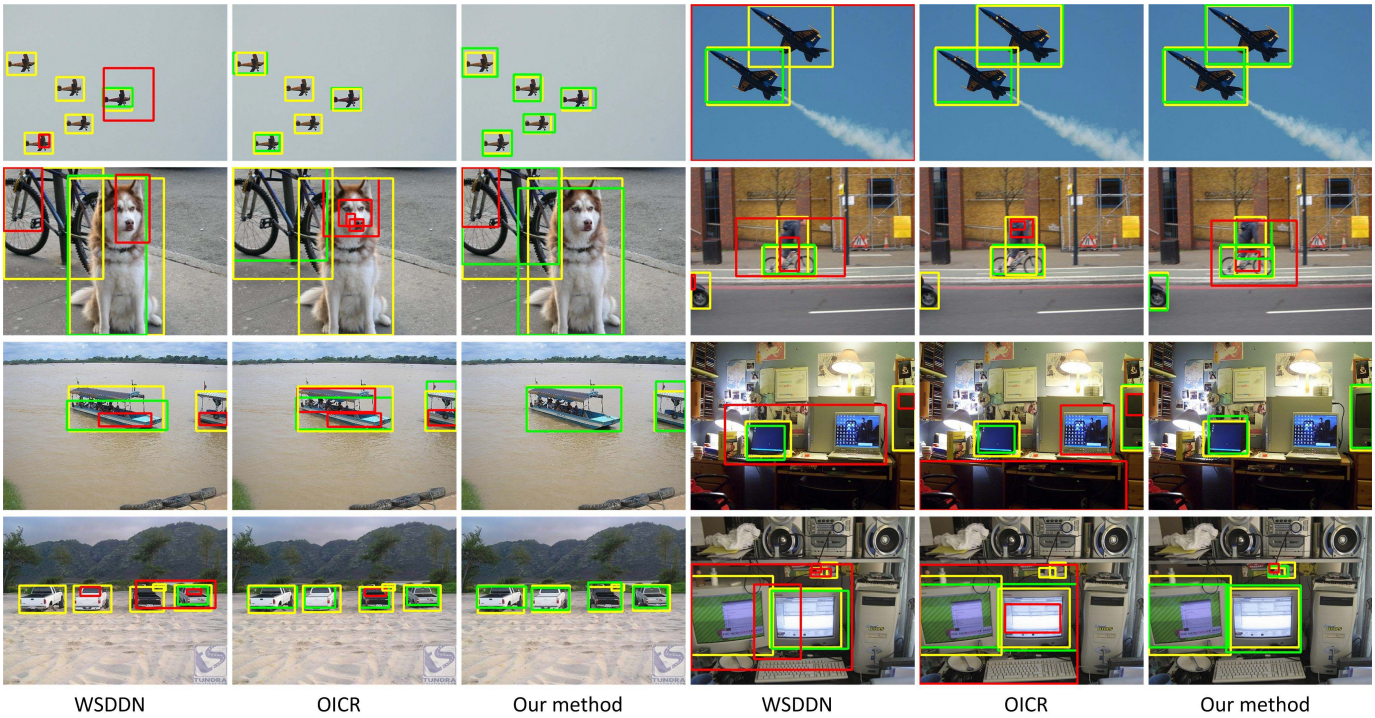


Fig. 9. Qualitative comparisons among the WSDDN [1], OICR [2] and our proposed PG-PS method on the VOC 2007 test set. Yellow rectangles indicate ground-truth bounding boxes. The successful detections (IOU$>$0.5) are marked with green bounding boxes, and the failed ones (IOU$<$0.5) are marked with red bounding boxes.

Influence of PG module. Our proposed PG module introduces about 82 additional proposals per image compared to the original SS method [21]. Thus, by introducing more high-quality proposals through Grad-CAM brings about 3.8% mAP gain (45.0% vs. 41.2%). As a contrast, we use Edge Boxes [54] to replace our Grad-CAM based method by picking out 100 proposals with the top-100 highest predictive scores (EB100). As seen, our method outperforms EB100 by 1.8% (45.0% vs. 43.2%). Figure 8 shows the IOU distribution of the proposals generated by using three different methods on the PASCAL VOC 2007 dataset. The results suggest that our PG module provides more proposals with high IOU.

Besides, we also select another weakly supervised object localization method, named Class Activation Mapping (CAM) [10], for comparison with Grad-CAM. As can be seen from Table I, CAM obtains slightly low mAP against Grad-CAM (44.6% vs. 45.0%). Actually, the results are not difficult to understand because Grad-CAM is a generalization of CAM and it often has better performance than CAM because Grad-CAM introduces a new scheme for combining feature maps to avoid any modification of the network, but CAM needs to modify the CNN architectures through replacing the fully-connected layers with convolutional layers and global average pooling operation, thus having an accuracy trade-off.

In addition, in order to further investigate the effectiveness of our coarse-to-fine, two-stage strategy for generating object proposals, we conduct ablation studies by using the first-stage coarse classifiers and the second-stage fine classifiers, respectively. Table II reports the results. As can be seen, the second-stage classifiers can significantly improve the detection accuracy from 42.7% to 45.0%. This is consistent with the class-specific activation maps as shown in Figure 4.

Influence of PS module. Our PS module can significantly improve the mAP of baseline method OICR [2]. To be specific,

TABLE I

ABLATION STUDIES BY USING THE BASELINE METHOD OICR+SS [2] AND ITS COMBINATION WITH VARIOUS COMPONENTS INCLUDING PROPOSAL GENERATION (PG), PROPOSAL SELECTION (PS), AND UP-WEIGHTING SCHEME FOR HARD NEGATIVES

	PG				PS				PG+PS
OICR+SS [2]	✓	✓	✓	✓	✓	✓	✓	✓	✓
EB100		✓							
CAM			✓						
Grad-CAM				✓					✓
Positive PS					✓			✓	✓
Negative PS						✓	✓	✓	✓
Up-weighting							✓	✓	✓
mAP (%)	41.2	43.2	44.6	45.0	46.1	45.5	46.8	47.7	51.1

TABLE II

ABLATION STUDIES OF THE GRAD-CAM BASED PROPOSAL GENERATION METHOD BY USING THE FIRST-STAGE CLASSIFIERS AND THE SECOND-STAGE CLASSIFIERS

	Grad-CAM based proposal generation	
The first-stage classifiers	✓	
The second-stage classifiers		✓
mAP (%)	42.7	45.0

TABLE III
DETECTION AVERAGE PRECISION (%) ON THE PASCAL VOC 2007 TEST SET

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN [1]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
DSTL [55]	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
WCCN [56]	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 [57]	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 [43]	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
MELM [3]	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
ZLDN [58]	55.4	68.5	50.1	16.8	20.8	62.7	66.8	56.5	2.1	57.8	47.5	40.1	69.7	68.2	21.6	27.2	53.4	56.1	52.5	58.2	47.6
C-WSL [59]	62.7	63.7	40.0	25.5	17.7	70.1	68.3	38.9	25.4	54.5	41.6	29.9	37.9	64.2	11.3	27.4	49.3	54.7	61.4	67.4	45.6
WSRPN [60]	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
ML-LocNet [61]	60.8	70.6	47.8	30.2	24.8	64.9	68.4	57.9	11.0	51.3	55.5	48.1	68.7	69.5	28.3	25.2	51.3	56.5	60.0	43.1	49.7
W2F+FRCNN [39]	63.5	70.1	50.5	31.9	14.4	72.0	67.8	73.7	23.3	53.4	49.4	65.9	57.2	67.2	27.6	23.8	51.8	58.7	64.0	62.3	52.4
C-MIL (VGG16) [36]	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
Pred Net (VGG16) [40]	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
OICR+GAM+REG [37]	55.2	66.5	40.1	31.1	16.9	69.8	64.3	67.8	27.8	52.9	47.0	33.0	60.8	64.4	13.8	26.0	44.0	55.7	68.9	65.5	48.6
WS-JDS+FRCNN [41]	64.8	70.7	51.5	25.1	29.0	74.1	69.7	69.6	12.7	69.5	43.9	54.9	39.3	71.3	32.6	29.8	57.0	61.0	66.6	57.4	52.5
OICR (baseline) [2]	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
OICR+FRCNN [2]	65.5	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	63.1	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
PG-PS (ours)	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
PG-PS+FRCNN (ours)	59.3	66.2	55.4	35.2	22.3	69.7	70.2	73.8	29.4	63.6	47.9	78.1	67.9	68.2	12.2	24.9	43.2	63.7	73.2	66.8	**54.6**

TABLE IV
LOCALIZATION PRECISION (%) ON THE PASCAL VOC 2007 TRAINVAL SET MEASURED IN TERMS OF CORLOC [50] METRIC

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN [1]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
DSTL [55]	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
WCCN [56]	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
PCL [57]	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 [43]	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
MELM [3]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.4
ZLDN [58]	74.0	77.8	65.2	37.0	46.7	75.8	83.7	58.8	17.5	73.1	49.0	51.3	76.7	87.4	30.6	47.8	75.0	62.5	64.8	68.8	61.2
C-WSL [59]	86.3	80.4	58.3	50.0	36.6	85.8	86.2	47.1	42.7	81.5	42.2	42.6	50.7	90.0	14.3	61.9	85.6	64.2	77.2	82.4	63.3
WSRPN [60]	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
ML-LocNet [61]	81.7	82.9	68.7	44.4	53.9	80.3	88.9	70.5	32.6	74.0	62.7	61.7	81.4	91.6	46.0	60.6	75.2	69.2	78.7	65.8	68.6
W2F+FRCNN [39]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	70.3
C-MIL (VGG16) [36]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.0
Pred Net (VGG16) [40]	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
OICR+GAM+REG [37]	81.7	81.2	58.9	54.3	37.8	83.2	86.2	77.0	42.1	83.6	51.3	44.9	78.2	90.8	20.5	56.8	74.2	66.1	81.0	86.0	66.8
WS-JDS+FRCNN [41]	79.8	84.0	68.3	40.2	61.5	80.5	85.8	75.8	29.7	77.7	49.5	67.4	58.6	87.4	66.2	46.6	78.5	73.7	84.5	72.8	68.6
OICR (baseline) [2]	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
OICR+FRCNN [2]	85.8	82.7	62.8	45.2	43.5	84.8	87.0	46.8	15.7	82.2	51.0	45.6	83.7	91.2	22.2	59.7	75.3	65.1	76.8	78.1	64.3
PG-PS (ours)	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
PG-PS+FRCNN (ours)	87.1	84.4	70.6	57.7	46.1	85.7	88.1	85.6	46.7	87.2	45.9	83.4	85.6	90.1	18.1	59.7	82.4	68.2	85.3	86.1	**72.2**

our positive PS and negative PS introduce 4.9% (from 41.2% to 46.1%) and 4.3% (from 41.2% to 45.5%) mAP gains, respectively. This demonstrates that through selecting as many positives as possible and mining discriminative hard negatives can help boost the performance significantly.

Influence of up-weighting scheme. To investigate the effect of up-weighting scheme, as shown in Eq. 9, we treat all mined negative proposals with the same weight, i.e., without incorporating Eq. 9 into Eq. 10. In this way, we obtain 45.5% mAP by just using negative PS. By using our proposed up-weighting scheme can boost mAP by 1.3% from 45.5% to 46.8%. In addition, by combining PG, PS and up-weighting scheme, we can largely improve OICR [2] by 9.9% mAP, which surpasses most of the state of the art WSOD methods.

D. Experimental Results on the VOC 2007 Dataset

We denote our approach as PG-PS for short. Besides, similar to most WSOD methods [2], [39], [57], [60]–[62], we also train a Faster RCNN [14] (FRCNN) detector by using the top-scoring proposals obtained with our PG-PS method as ground truths (PG-PS+FRCNN in tables), which can further significantly improve the performance.

A total of 16 state-of-the-art WSOD methods are chosen for comparison and most of them are published in the last two years. They are WSDDN [1], OICR [2], DSTL [55], WCCN [56], PCL [57], TS2C [43], MELM [3], ZLDN [58], C-WSL [59], WSRPN [60], ML-LocNet [61], W2F [39], C-MIL [36], Pred Net [40], OICR+GAM+REG [37], and WS-JDS [41].

The detailed result comparisons among our method and other methods in terms of mAP on the PASCAL VOC 2007 test set and CorLoc on the VOC 2007 trainval set are shown in Table III and Table IV, respectively. As we can see, our proposed PG-PS+FRCNN method outperforms all comparison methods by using both mAP and CorLoc measurements. Especially, compared to the OICR method [2] which is the baseline of our method, we obtain 13.4% mAP gain and 11.6% average CorLoc gain, which are significant big margins for the challenging WSOD task. Compared with previous best-performing method Pred Net (VGG16) [40]

TABLE V

DETECTION AVERAGE PRECISION (%) ON THE PASCAL VOC 2012 TEST SET. RESULTS: †http://host.robots.ox.ac.uk:8080/anonymous/EAFZZP.html, ‡http://host.robots.ox.ac.uk:8080/anonymous/HNKB4R.html

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DSTL [55]	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
WCCN [56]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.9
PCL [57]	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 [43]	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
MELM [3]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	42.4
ZLDN [58]	54.3	63.7	43.1	16.9	21.5	57.8	60.4	50.9	1.2	51.5	44.4	36.6	63.6	59.3	12.8	25.6	47.8	47.2	48.9	50.6	42.9
WSRPN [60]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.8
ML-LocNet [61]	53.9	60.4	40.4	23.3	18.7	58.7	63.3	52.5	13.3	49.1	46.8	33.5	61.0	65.8	21.3	22.9	46.8	48.1	52.6	40.4	43.6
W2F+FRCNN [39]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	47.8
C-MIL (VGG16) [36]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.7
Pred Net (VGG16) [40]	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
OICR+GAM+REG [37]	64.7	66.3	46.8	28.5	28.4	59.8	58.6	70.9	13.8	55.0	15.7	60.5	63.9	69.2	8.7	23.8	44.7	52.7	41.5	62.6	46.8
WS-JDS+FRCNN [41]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.1
OICR (baseline) [2]	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
OICR+FRCNN [2]	71.4	69.4	55.1	29.8	28.1	55.0	57.9	24.4	17.2	59.1	21.8	26.6	57.8	71.3	1.0	23.1	52.7	37.5	33.5	56.6	42.5
PG-PS (ours)†	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
PG-PS+FRCNN (ours)‡	70.8	69.9	51.9	27.3	28.1	65.2	62.1	82.4	24.2	55.8	30.6	79.7	64.4	72.5	3.2	27.1	45.6	61.9	69.7	63.7	**52.9**

TABLE VI

LOCALIZATION PRECISION (%) ON THE PASCAL VOC 2012 TRAINVAL SET IN TERMS OF CORLOC [50] METRIC

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DSTL [55]	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
PCL [57]	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 [43]	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
ZLDN [58]	80.3	76.5	64.2	40.9	46.7	78.0	84.3	57.6	21.1	69.5	28.0	46.8	70.7	89.4	41.9	54.7	76.3	61.1	76.3	65.2	61.5
WSRPN [60]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.9
ML-LocNet [61]	88.1	85.5	71.2	49.4	57.4	90.7	77.6	53.5	42.6	79.6	34.1	69.1	81.7	91.9	35.4	64.6	79.3	64.3	79.3	69.6	68.2
W2F+FRCNN [39]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.4
C-MIL (VGG16) [36]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.4
Pred Net (VGG16) [40]	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
OICR+GAM+REG [37]	82.4	83.7	72.4	57.9	52.9	86.5	78.2	78.6	40.1	86.4	37.9	67.9	87.6	90.5	25.6	53.9	85.0	71.9	66.2	84.7	69.5
WS-JDS+FRCNN [41]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.5
OICR (baseline) [2]	86.2	84.2	68.7	55.4	46.5	82.8	74.9	32.2	46.7	82.8	42.9	41.0	68.1	89.6	9.2	53.9	81.0	52.9	59.5	83.2	62.1
OICR+FRCNN [2]	89.3	86.3	75.2	57.9	53.5	84.0	79.5	35.2	47.2	87.4	43.4	43.8	77.0	91.0	10.4	60.7	86.8	55.7	62.0	84.7	65.6
PG-PS (ours)	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
PG-PS+FRCNN (ours)	85.8	83.1	73.4	54.1	43.5	87.9	82.2	85.5	49.3	83.4	46.4	81.3	86.5	87.9	16.6	53.7	80.2	74.1	83.5	81.7	**71.0**

(52.9% mAP and 70.9% CorLoc), our proposed method (54.6% mAP and 72.2% CorLoc) still achieves 1.7% mAP and 1.3% CorLoc gains.

In Figure 9, we visualize some detection results on the PASCAL VOC 2007 test set. The results from WSDDN [1] and OICR [2] are employed for comparison. In the figures, yellow rectangles indicate ground-truth bounding boxes, green bounding boxes denote successful detections (IOU$>$0.5), and red bounding boxes represent failed ones (IOU$<$0.5). As can be seen from Figure 9, our proposed approach can produce much tighter bounding boxes and achieve more precise localization, whereas other two methods always fail to generate boxes that are either overlarge or only contain parts of the objects. In particular, when multiple objects from the same category appear in an image, our method can accurately detect them with bigger IOU, but other two methods usually have some missed detections. These results strongly validate our motivation.

E. Experimental Results on the VOC 2012 Dataset

Table V and Table VI show the detection average precision and localization results of our approach and the state-of-the-art methods on the PASCAL VOC 2012 dataset, respectively. Our method improves the state-of-the-arts to 52.9% and 71.0%, measured in terms of mAP and average CorLoc, respectively, both of which are the best results. Moreover, the results greatly improve the baseline method OICR [2] by large margins: 15.0% mAP and 8.9% average CorLoc gains. The similar results to that on the VOC 2007 dataset demonstrate the effectiveness of our proposed method.

TABLE VII

DETECTION AVERAGE PRECISION (%) AND LOCALIZATION PRECISION (%) ON THE COCO DATASET

Method	mAP (%)	CorLoc (%)
WSDDN [1]	11.5	26.1
MELM [3]	18.8	-
ML-LocNet [61]	16.2	34.7
WS-JDS [41]	20.3	-
OICR [2] (baseline)	14.3	30.4
PG-PS (**ours**)	**20.7**	**35.4**

F. Experimental Results on the COCO Dataset

Table VII reports the detection results and localization results of our approach on the COCO dataset. Since only few works have reported the results on the COCO dataset under weakly supervised learning paradigm, we here only compare our method with five popular WSOD method including WSDDN [1], MELM [3], ML-LocNet [61], WS-JDS [41], and OICR [2]. As seen from Table VII, on this challenging dataset, our method improves the baseline method OICR [2] with 6.4% mAP gain and 5.0% CorLoc gain, obtaining state-of-the-art results for both detection accuracy and localization accuracy.

V. Conclusion

This paper proposed an effective method for weakly supervised object detection, which is achieved by adding two important modules to the state-of-the-art OICR system [2]. One is proposal generation, which focuses on generating more object proposals with higher IOU. The other one is proposal selection, which aims to choose as many positive proposals as possible and mine discriminative hard negatives to make training more effective. Experimental results on the PASCAL VOC 2007 and 2012 datasets and MS COCO dataset demonstrate that our method significantly improves the baseline method OICR [2] by large margins and achieves the state-of-the-art results compared with existing WSOD methods.

References

[1] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2846–2854.

[2] 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.

[3] F. Wan, P. Wei, Z. Han, J. Jiao, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2395–2409, Oct. 2019.

[4] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1325–1334.

[5] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "P-CNN: Part-based convolutional neural networks for fine-grained visual categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 6, 2019, doi: 10.1109/TPAMI.2019.2933510.

[6] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1277–1286.

[7] J. R. R. Uijlings, S. Popov, and V. Ferrari, "Revisiting knowledge transfer for training object class detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1101–1110.

[8] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang, "Generative adversarial learning towards fast weakly supervised detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5764–5773.

[9] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?–Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 685–694.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[11] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Two-phase learning for weakly supervised object localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3534–3543.

[12] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3544–3553.

[13] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

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

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[16] G. Cheng, P. Zhou, and J. Han, "RIFD-CNN: Rotation-invariant and Fisher discriminative convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2884–2893.

[17] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

[18] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[19] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[20] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[21] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[22] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. Manjunath, "Weakly supervised localization using deep feature maps," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 714–731.

[23] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.

[24] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Image co-localization by mimicking a good detectors confidence score distribution," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 19–34.

[25] M. Shi and V. Ferrari, "Weakly supervised object localization using size estimates," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 105–121.

[26] W. Ren, K. Huang, D. Tao, and T. Tan, "Weakly supervised large scale object localization with multiple instance learning and bag splitting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 405–416, Feb. 2016.

[27] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3512–3520.

[28] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari, "We Don't need no bounding-boxes: Training object class detectors using only human verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 854–863.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[30] M. Shi, H. Caesar, and V. Ferrari, "Weakly supervised object localization using things and stuff transfer," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3381–3390.

[31] X. Wang, Z. Zhu, C. Yao, and X. Bai, "Relaxed multiple-instance SVM with application to object discovery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1224–1232.

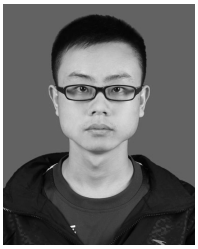
[32] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold MIL training for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2409–2416.

[33] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko, "Detector discovery in the wild: Joint multiple instance and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2883–2891.

[34] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3460–3469.
[35] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1081–1089.
[36] 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.
[37] 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.
[38] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "ContextlocNet: Context-aware deep network models for weakly supervised localization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 350–365.
[39] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem, "W2F: A weakly-supervised to fully-supervised framework for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 928–936.
[40] 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.
[41] 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.
[42] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 73–80.
[43] Y. Wei *et al.*, "Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 434–450.
[44] Z. Chen, S. Huang, and D. Tao, "Context refinement for object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 71–86.
[45] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Soft proposal networks for weakly supervised object localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1841–1850.
[46] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang, "Self-produced guidance for weakly-supervised object localization," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 597–613.
[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.
[48] 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.
[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.
[50] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 275–293, Dec. 2012.
[51] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.
[52] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.
[53] 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.
[54] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 391–405.
[55] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep self-taught learning for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1377–1385.
[56] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 914–922.
[57] 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.
[58] 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.
[59] 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.*, Sep. 2018, pp. 152–168.
[60] P. Tang *et al.*, "Weakly supervised region proposal network and object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 352–368.
[61] X. Zhang, Y. Yang, and J. Feng, "Ml-LocNet: Improving object localization with multi-view learning network," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 240–255.
[62] J. Son, D. Kim, S. Lee, S. Kwak, M. Cho, and B. Han, "Forget & diversify: Regularized refinement for weakly supervised object detection," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 632–648.

Gong Cheng received the B.S. degree from Xidian University, Xi'an, China, in 2007, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, in 2010 and 2013, respectively. He is currently a Professor with Northwestern Polytechnical University. His main research interests are computer vision, pattern recognition, and remote sensing image understanding. He is an Associate Editor of the *IEEE Geoscience and Remote Sensing Magazine* and a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Junyu Yang received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2016, where he is currently pursuing the master's degree. His main research interests are computer vision and pattern recognition.



Decheng Gao received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2017, where he is currently pursuing the master's degree. His main research interests are computer vision and pattern recognition.

Lei Guo received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 1982 and 1986, respectively, and the Ph.D. degree from Northwestern Polytechnical University, Xi'an, in 1993. He is a Professor with the School of Automation, Northwestern Polytechnical University. His research interest focuses on image processing.

Junwei Han (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in pattern recognition and intelligent systems from Northwestern Polytechnical University, Xi'an, China, in 1999, 2001, and 2003, respectively. He was a Research Fellow with Nanyang Technological University, The Chinese University of Hong Kong, Dublin City University, and the University of Dundee from 2003 to 2010. He is currently a Professor with Northwestern Polytechnical University. His research interests include computer vision and brain imaging analysis. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON MULTIMEDIA.