Progressive Contextual Instance Refinement for Weakly Supervised Object Detection in Remote Sensing Images

Xiaoxu Feng, Junwei Han, *Senior Member, IEEE*, Xiwen Yao, and Gong Cheng, *Member, IEEE*

***Abstract*— Weakly supervised learning has been attracting much attention due to its broad applications, which only requires image-level annotations to indicate whether there exist objects in the images. Currently, most of the existing weakly supervised object detection (WSOD) methods are inclined to seek only one top-scoring object instance per image from noisy proposals to train the corresponding object detector. However, more than one same-class instances often exist in the large-scale, cluttered remote sensing images. Thus, selecting only one top-scoring proposal usually results in highlighting the most representative part of an object rather than the whole object, which may cause learning a suboptimal object detector by losing much important information. To address this problem, a novel end-to-end progressive contextual instance refinement (PCIR) method is proposed to perform WSOD. Specifically, a dual-contextual instance refinement (DCIR) strategy is designed to divert the focus of the detection network from the local distinct part to the whole object and further to other potential instances by leveraging both local and global context information. Benefiting from DCIR, a progressive proposal self-pruning (PPSP) strategy is further developed to mitigate the influence of the complex background by dynamically rejecting the negative training proposals. Comprehensive experiments on the challenging NWPU VHR-10.v2 and DIOR data sets clearly demonstrate that the proposed method can significantly boost the detection accuracy compared with the state of the arts.**

***Index Terms*— Contextual instances refinement, weakly supervised object detection (WSOD).**

NOMENCLATURE

$I^{(i)}$	Input image.
Y	Image-level label.
R	Proposals.
C	Number of object classes.
$\|R\|$	Number of proposals.
x_{ij}^{c}	Proposal's feature vectors in the classification stream.
x_{ij}^{d}	Proposal's feature vectors in the detection stream.
x_{cr}	Proposal's score belonging to class c.
P_c	Image-level score for the image belonging to class c.
x_{cr}^{k}	Proposal's score in the kth refinement branch.
y_{cr}^{k}	Pseudo label in the kth refinement branch.
R_{c1}^{n}	nth candidate to class c.
R_{cn}	nth candidate's contextual regions corresponding to class c.
V_{cn}	nth unified contextual representation to class c.
$S_{V_{cn}}$	New confidence of the nth candidate to class c.
s	sth training stage.
$x_{cR_1}^{(k-1)}$	Top-scoring in the $\{k-1\}$th refinement branch.
P_{sp}^{n}	Proposals after n times of pruning stage.

Manuscript received March 10, 2020; accepted April 1, 2020. Date of publication April 27, 2020; date of current version October 27, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0502900, in part by the National Science Foundation of China under Grant 61701415, Grant 61772425, and Grant 61773315, in part by the Fundamental Research Funds for Central Universities under Grant 3102019ZDHKY05, in part by China Postdoctoral Science Foundation under Grant 2018T111094 and Grant 2017M620468, and in part by the Postdoctoral Science Foundation of Shaanxi Province under Grant 2017BSHYDZZ36. *(Corresponding authors: Junwei Han; Xiwen Yao.)*

Xiaoxu Feng, Junwei Han, and Gong Cheng are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

Xiwen Yao is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Qingdao Research Institute, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yaoxiwen@nwpu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.2985989

I. INTRODUCTION

WITH the development of modern remote sensor technologies, object detection from remote sensing images has become an indispensable task for various aerial and satellite image applications [1]–[6], such as traffic planning, remote sensing retrieval, and urban construction. Recently, benefiting from the stronger feature representation power of convolutional neural networks (CNNs) [7], [8] and the availability of abundant data sets with subtle instance-level annotations [9], [10], object detection has achieved breakthrough performance. However, collecting such precise annotations is labor-intensive and time-consuming. In contrast, image-level labels can be easily obtained. In this article, we focus on applying weakly supervised learning techniques to achieve the task of object detection in remote sensing images.

Most methods [11]–[15] address the weakly supervised object detection (WSOD) problem by treating each image as a bag of potential object instances and then training instance classifiers under multiple instance learning (MIL) constraints. Based on this, many efforts [16]–[19] have been made to take a two-stage approach to obtain better WSOD performance. First, object proposal extraction methods [20], [21] are used to generate a set of candidate boxes. Then, object detection is treated as the instance-level classification problem after

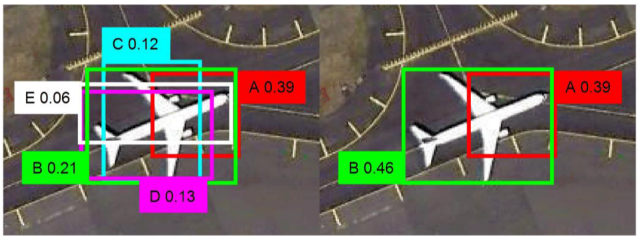


Fig. 1. Detection results with (Left) OICR method and (Right) our PICR. In the Left, proposal A with the top score of 0.39 does not correctly cover the whole object. By using our strategy, as shown in the Right, the correct proposal B is accordingly highlighted by increasing its confidence score from 0.21 to 0.46.

extracting the CNN features of each proposal. Another line of work [22] focuses on mining easily discriminating and high confident examples as the pseudo ground truths, and then, WSOD is formulated as a fully supervised problem.

Although promising results have been reported in the aforementioned methods [11]–[19], [22], there are still two big challenges for the development of WSOD in remote sensing images. The first challenge is that most of WSOD approaches incline to only select the most confident candidate box to train the corresponding detector. However, the top-scoring proposal often covers only one part of the object or even confusing backgrounds rather than the entire object, particularly in large-scale, cluttered remote sensing images. Hence, the aforementioned approaches may be insufficient to automatically detect the objects in satellite images. Moreover, there usually are more than one same-class instances in remote sensing images. Simply selecting one proposal per class as the pseudo ground truth in an image leads to suboptimal object detectors and results in an inferior performance of WSOD compared with those fully supervised approaches.

The second challenge is the ambiguity for learning discriminative object detection models caused by the absence of instance-level supervision. What is worse is that the large-scale, cluttered backgrounds in remote sensing images offer more negative samples, which further increases the learning difficulty. Although many proposal generation algorithms [20], [21] have attempted to alleviate this problem by reducing the imbalance proportion of foreground and background samples, the influences of negative samples have not been effectively solved.

To address the first challenge, we elaborately propose an end-to-end dual-contextual instance refinement (DCIR) strategy to divert the focus of the detection network from the local distinct part to the whole object and further to other potential instances by leveraging both local and global context information. More specifically, instead of selecting the proposal with the highest score as the pseudo supervision, we first select a confident candidate box from cluttered proposals and identify its neighboring regions. Next, the local contextual information from the neighboring regions is gathered to aggregate into a unified contextual representation. The new confidence will be generated according to the unified contextual representation. As shown in Fig. 1, proposal A is assigned with the highest score of 0.39 by the method of online instance classifier refinement (OICR) [17]. Unfortunately, the proposal A only covers a part of the airplane. In contrast, our method highlights proposal B with its score increased from 0.21 to 0.46 and can successfully cover the whole object. As the refinement proceeds, the proposal that contains the whole object instead of an object part is assigned with the top score. After that, we greedily and iteratively propagate the label information from the most confident local part to the global suboptimal regions that have similar confidences but without spatial connections. Finally, we apply these pseudo instance-level labels to train an end-to-end instance detector.

To address the second challenge, a progressive proposal self-pruning (PPSP) strategy is designed to alleviate the influence of negative samples introduced by complex backgrounds. Specifically, we iteratively reject well-classified negative proposals. Considering the poor performance of the network at the beginning, it is inevitable to reject the positive instances. On the one hand, the proposals will be pruned with the improvement of network performance. On the other hand, the proposals will be reset at regular intervals to prevent losing true positive regions. Accordingly, it can not only alleviate the influence of negative examples for classification but also maintain the diversity of proposals to prevent the model trapping in a local optimum.

Combining DCIR with PPSP formulates a progressive contextual instance refinement (PCIR) process. The DCIR module endeavors to mine all potential instances existing in the images, while the PPSP module removes noisy negative samples and maintains the diversity of the samples. Experiments on the NWPU VHR-10.v2 and DIOR data sets clearly demonstrate the effectiveness of our method compared with the state-of-the-art methods under a weakly supervised paradigm. To sum up, our main contributions can be summarized as follows.

1) We propose a novel DCIR algorithm that leverages the surrounding context information to suppress low-quality object parts and further to highlight the whole objects with high quality. Moreover, we adopt a contextual confidence similarity discovery strategy to mine all possible instances existing in the images to train the object detector.
2) We introduce a PPSP algorithm during the end-to-end learning process, which can filter out confusing negative samples and meanwhile maintain the diversity of training samples.
3) Experimental results demonstrate that the proposed method achieves state-of-the-art results and comparable performance with many classic fully supervised object detection methods.

II. Related Work

A. Object Detection in Remote Sensing Images

In the past decades, object detection in remote sensing images has been extensively studied. Before CNN [7], [8] was widely used, many approaches [23]–[29] treated object detection in remote sensing images as a classification problem. The main purpose of these methods is to extract more discriminative features. Among them, one of the most popular methods is the bag-of-words (BoW) method [23]. The BoW model can effectively achieve the task of object detection by

quantizing each image region into a visual word and then calculating a histogram representation. In addition, the histogram of oriented gradients (HOG) [24] used the distribution of gradient strength and the direction in the spatial distribution area to represent the objects. On the basis of the HOG feature, the sparselets work [25] has been successfully applied to remote sensing image analysis. Then, the development of compressed sensing promotes many methods [26]–[29] applying sparse coding to the task of object detection in remote sensing images.

After that, many deep learning-based methods [10], [30]–[34] addressed object detection problem in a fully supervised manner and achieved significant improvements. For instance, Cheng *et al.* [30] proposed a rotation-invariant CNN (RICNN) model that can tackle the variations of object orientation in remote sensing images by designing a novel rotation-invariant layer. Tang *et al.* [31] successfully achieved the task of detecting vehicles in remote sensing images by establishing a hyperregion proposal network (HRPN) and a cascade of boosted classifiers. Yang *et al.* [32] built the Markov random field (MRF)-fully convolutional network to detect airplanes.

Although many fully supervised learning approaches have achieved good performance, they need precise instance-level annotations. It is labor-intensive and time-consuming. Thus, several efforts were made to address the problem of object detection in remote sensing images under a weakly supervised learning paradigm. For example, Han *et al.* [35] iteratively trained the object detector by using refined annotations until the model converges. Zhou *et al.* [36] proposed a novel weakly object detection method based on negative bootstrapping and transferred deep features in remote sensing images.

Recently, many popular methods [11], [12], [14], [16], [18], [22] apply MIL to address the WSOD problem and achieve promising results. However, they cannot be directly applied to address the problem of object detection in remote sensing images.

B. Context Refinement in Object Detection

Recently, there have been comprehensive studies that applied visual context to improve the object detection performance. Many studies [37]–[43] have demonstrated that using context can effectively boost the performance of object detection even without the powerful CNN. More recently, many methods further incorporate contexts with CNN to facilitate detection. For instance, Bell *et al.* [44] introduced a multidirectional recurrent neural network to gather context. Wei *et al.* [45] leveraged segmentation context to mine tight boxes. Li *et al.* [10] employed additional multiangle anchors with region proposal network (RPN) and double-channel feature fusion network to handle the problem of multiangle, multiscale and appearance ambiguity in remote sensing images.

III. Proposed Method

A. Notations

Before presenting our method, we first introduce the notations used in this article, as shown in the Nomenclature.

B. Overview of the Proposed Method

The overall architecture of the proposed WSOD framework is illustrated in Fig. 2. The core goal of our method is to mine more accurate positive instances through taking full advantage of the context information under weakly supervised settings.

To this end, a novel PCIR method is designed, which mainly consists of two modules, namely, DCIR and PPSP. Specifically, for each input image $I^{(i)}$, the selective search [20] method is used to generate about 2000 proposals, and the region of interest (RoI) pooling [46] is employed to generate fixed-size convolution features. Following two fully connected layers, the proposal features are branched into two streams: the upper stream performs classification for each individual proposal and the down stream performs instead detection. The score of each proposal is produced by elementwise production between the results of these two streams

$$x_{cr} = \frac{e^{x_{ij}^c}}{\sum_{k=1}^{C} e^{x_{kj}^c}} \odot \frac{e^{x_{ij}^d}}{\sum_{r=1}^{|R|} e^{x_{ir}^d}} \tag{1}$$

where C denotes the number of image categories and $|R|$ denotes the number of proposals. Finally, the image-level score of the Cth class can be generated by the summation over all proposals

$$P_c = \sum_{r=1}^{|R|} x_{cr}. \tag{2}$$

The loss function [11] of the two-stream network can be formulated as follows:

$$\text{Loss}_{\text{wsddn}} = -\sum_{c=1}^{C} \{Y \log P_c + (1 - Y) \log(1 - P_c)\}. \tag{3}$$

In parallel with the two-stream network, several refinement branches are added. Each refinement branch maps the proposal's feature vector to a $\{C + 1\}$-dimensional score vector $x_{cr}^k \in \mathbb{R}^{(C+1)\times 1}, k \in (1, \ldots, K)$, where k represents the kth refinement branch and $\{C + 1\}$ represents C different object classes and one background class.

In each refinement branch, the proposed DCIR strategy is introduced to obtain the most representative whole object by using the local context information and further to discover other possible instances existing in the images by using the global context information. Denote $y_{cr}^k \in (0, \ldots, C)$ as the pseudo label of the rth proposal in the kth refinement branch, and the loss function [17] for the kth refinement is defined as

$$\text{Loss}_r^k = -\frac{1}{|R|} \sum_{r=1}^{|R|} \sum_{c=1}^{C+1} y_{cr}^k \log x_{cr}^k. \tag{4}$$

After performing several refinements, we can obtain the pseudo labels of all potential instances. Then, these instances together with their pseudo labels are incorporated into the following PPSP module to filter out negative proposals and, meanwhile, maintain the diversity of training samples. Finally, we train the detection network by combining the loss functions of $\text{Loss}_{\text{wsddn}}$ and Loss_r^k

$$\text{Loss} = \text{Loss}_{\text{wsddn}} + \sum_{k=1}^{K} \text{Loss}_r^k \tag{5}$$

where K denotes the total number of the refinement times.

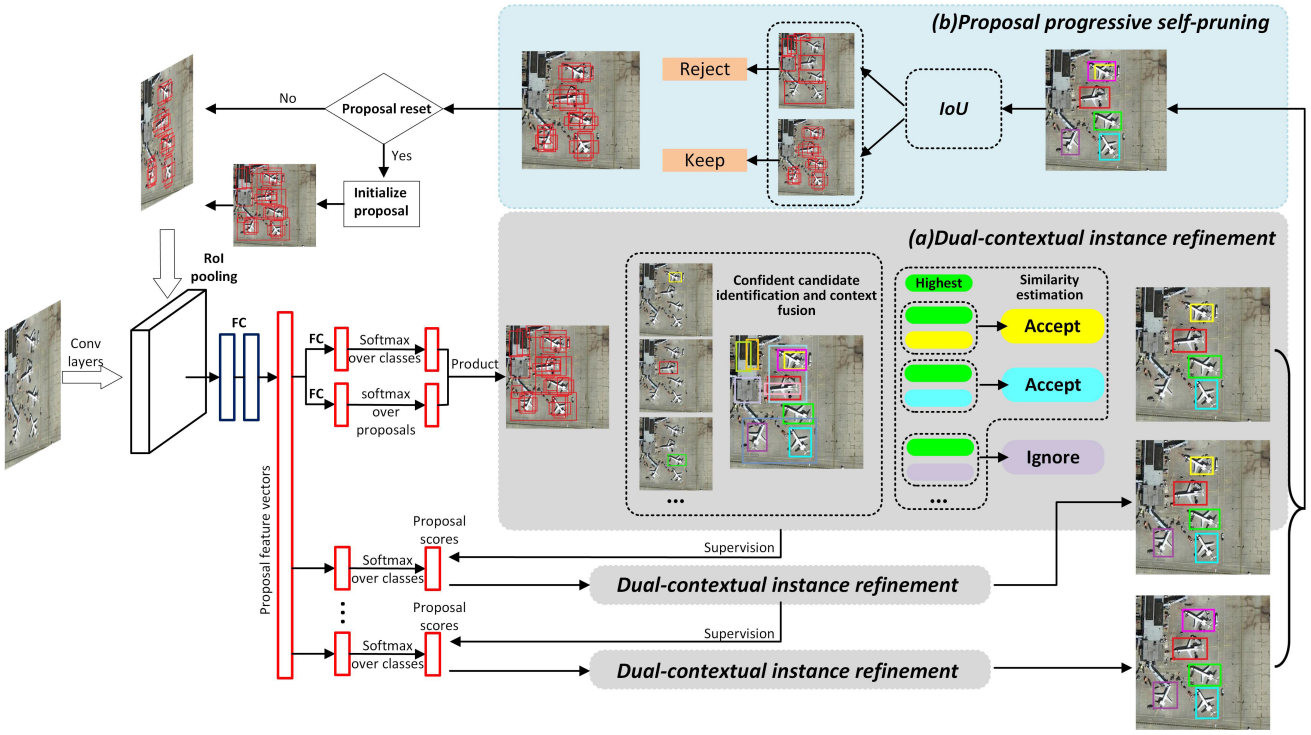


Fig. 2. Overview of the proposed PCIR method. (a) DCIR mines tight bounding box by fusing its surrounding regions' contexts and retrieves each instance in the images by estimating the context similarity between the highest-scored candidate and other proposals. (b) Proposal progress self-pruning strategy dynamically rejects negative proposals by calculating the IoU between each proposal and the positive examples generated by DCIR.

C. Dual-Contextual Instance Refinement

In this section, we will expound how to mine credible instances in the images when only image-level labels are available. First, we incorporate local context supplemented by neighboring regions to divert the focus of the detection network from the local distinct part to the whole object. After that, the relationship between local and global context information will be analyzed to propagate the label information to other possible instances existing in the images. These two stages formulate the DCIR process.

1) Confident Candidate Identification and Context Fusion: Give an image $I^{(i)}$ with the image-level label $Y = [y_1, \ldots y_c, \ldots, y_C]^T \in (0,1)^C, c \in \{1, \ldots, C\}$, we sort the proposals' scores $x_{cr}(r \in \{1, \ldots, R\})$ in descending order for each $y_c = 1$. First, a set of high-scoring proposals corresponding to class c is selected as the candidate of contextual representation R^n_{c1}. Inspired by OICR [17], when a candidate is selected, we identify the regions that have high spatial overlap with it as context regions.

To take advantage of local context information in the context region, the concept of correlation coefficient is introduced to describe the closeness between the candidate and other proposals. We use $\sigma(R^n_{c1}, R)$ to denote the correlation between candidate R^n_{c1} and proposal R. Using this notation, similar to [43], the composition of the nth candidate's contextual regions corresponding to class c can be expressed as

$$R_{cn} = \{R^n_{c1} | \sigma(R^n_{c1}, R) > \lambda\}, \quad j \neq 1 \tag{6}$$

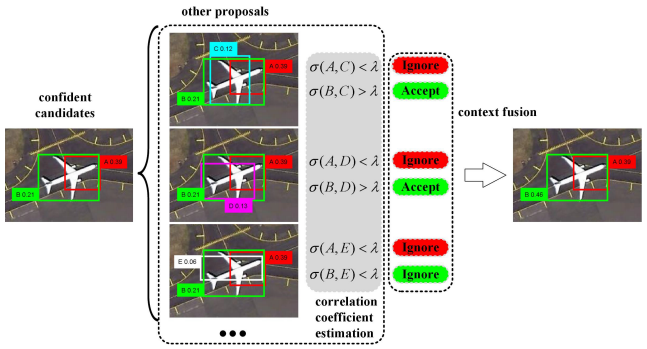


Fig. 3. Illustration of the process of context fusion in DCIR. It aims to highlight the high-quality proposal by using its surrounding regions' contexts.

where λ is a threshold. In our implementation, the correlation coefficient is measured by calculating the intersect-over-union (IoU) between two regions. Fig. 3 illustrates a concrete example of our candidate identification and context fusion.

Next, the nth candidate and its collected context regions are constructed to form a unified contextual representation V_{cn}. In order to prevent selecting the next candidate instance with close confidence score from the same location, the proposal that has the highest score in the remaining candidate boxes but without high spatial overlap with other candidates is selected as next candidate. Thus, the composition of the nth unified contextual representation is

$$\begin{cases} V_{cn} = \left\{ R^n_{c1} | R^n_{c1} \bigcup R_{cn} \right\} \\ \forall i \neq j, \quad R^i_{c1} \notin V_{cj}. \end{cases} \tag{7}$$

The new confidence of the nth candidate $S_{V_{cn}}$ is regenerated by fusing all the scores of the unified contextual representation.

After fusing the candidate and its local surrounding contexts, the nonmaxima suppression (NMS) [47] is adopted to remove redundant candidates so that the high-quality whole object is accordingly highlighted.

2) Similarity Estimation: As we know, the proposals with similar characteristics would have close detection scores. We assume that different instances belonging to the same object category should have close detection scores. Accordingly, the label information of local optimal candidate instance is propagated to global credible candidates by analyzing the contextual confidence similarity from a global perspective. To this end, the concept of credibility level is introduced to evaluate the similarity between the highest-scoring candidate and other candidates. Specifically, the highest-scoring candidate in each image is identified as a criterion for evaluating the credibility level. Thus, the credibility level of the ith candidate can be formulated as

$$\begin{cases} \mathrm{Con}_{V_{cn}} = {}^{S_{V_{cn}}}/_{ST_c}, & n \in [1, N],\ c \in [1, C] \\ ST_c = \max\{S_{V_{c1}}, \ldots, S_{V_{cn}}\}, & n \in [1, N],\ c \in [1, C] \end{cases} \tag{8}$$

where C represents the total number of object classes and N denotes the number of contextual representations. To mine each possible instance existing in an image, the dynamic adaptive growth threshold is applied to select all possible candidates. If an image has class label c, we assign the candidate with label c when the credibility level is larger than the threshold and 0 when the credibility level is lower than the threshold as follows:

$$y_{R_{cn}} = \begin{cases} c, & \mathrm{Con}_{V_{cn}} \geq T_c + k_c \times s \\ 0, & \mathrm{Con}_{V_{cn}} < T_c + k_c \times s \end{cases} \tag{9}$$

where k_c is the growth factor and $\mathrm{Con}_{V_{cn}}$ can be calculated by using (8). Here, we simply adopt the linear growth approach. Other approaches are possible, but this way works well in our experiments. As different proposals with high spatial overlap should share the same label information, we can label all the members of confident contextual representation to class c and label other proposals to class 0. The whole process of DCIR is summarized in Algorithm 1.

At the beginning, the instance-level supervised information is mixed with a lot of noise due to the lower confidence threshold. In order to reduce the impact of noise on the training, we change the loss function [17] in (4) as follows:

$$\mathrm{Loss}_r^k = -\frac{1}{|R|}\sum_{r=1}^{|R|}\sum_{c=1}^{C+1} w_r^k y_{cr}^k \log x_{cr}^{Rk},\ w_r^k = x_{cR_1}^{(k-1)}. \tag{10}$$

D. Progressive Proposal Self-Pruning

The instance-level pseudo labels discovered by the DCIR strategy provide good initialization for object detection, but the proposals generated in[20] and [21] are always noisy, which will introduce ambiguities for learning discriminative object detection models. Hence, the PPSP is designed to reduce the ambiguities for further improving the performance of multiple instance detection networks.

Algorithm 1 DCIR Algorithm

Input: Training data set $I = \{(I^{(1)}, Y^{(1)}), \ldots, (I^{(N)}, Y^{(N)})\}$ and its proposals; refinement times K.
Output: Loss weights w_r^{k+1}, Instance-level label $y_{R_{cn}}$

```
Feed image I^(i) and its proposals into the network to produce proposal score matrices x_cr^k, k ∈ {0, ..., K − 1}.
Set initial confidence threshold T_c for each class.
for k = 0 to K − 1 do
    for c = 1 to C do
        if y_c = 1 then
            Select candidates and identify their surrounding regions by Eq.(6) and Eq.(7).
            Generate new confidence by fusing context.
            Select the highest score as credibility level criterion ST_c and set loss weight w_r^{k+1} = x_{cR_1}^k
            Estimate the similarity by calculating Con_{V_cn} in Eq.(8)
            Label the contextual representations by Eq.(9)
        end if
    end for
end for
```

First, the contextual representations that are larger than the threshold are identified as positive instances. Next, the proposals are preliminarily pruned by rejecting well-classified negative proposals. Specifically, the IoU is calculated between each proposal and the credible candidates. The boxes with IoU larger than Pth are selected as new proposals for the next epoch. Thus, the training proposals in the nth pruning stage can be represented as

$$P_{\mathrm{sp}}^n = \{\mathrm{IoU}(P_j, R_{c1}^i) > P_{\mathrm{th}}\},\quad j \in \left\{1, \left|P_{\mathrm{sp}}^{n-1}\right|\right\} \tag{11}$$

where R_{c1}^i represents the ith credible candidate in class c and P_j denotes the proposals by performing the jth pruning. However, it is no negligible that the aforementioned pruning stage will reject the positive instances. If the pruning process has been maintained for a long time, the model will get trapped in suboptimal by losing much important information. Thence, the proposals will be reset to initial proposals at regular intervals. Accordingly, the collaboration between the pruning process and the proposal reset is termed a self-pruning stage. Although the instances may be rejected at the first self-pruning stage, the false-negative instances will be discovered by the next stage. With the performance improvement of the network, the proposals selected by the self-pruning strategy will become more and more accurate. Meantime, depending on the performance of the network, the candidates mined by DCIR are different so that the training proposals will be diverse. Following this self-pruning strategy, we can not only reject the noisy proposals but also maintain the diversity of the training samples.

At the beginning of training, due to the poor performance of the object detector, it is inevitable that some false examples will be labeled as pseudo instances. If we filter the proposals based on the pseudo labels at this stage, it will lead to the extreme phenomena that choosing the true positive examples

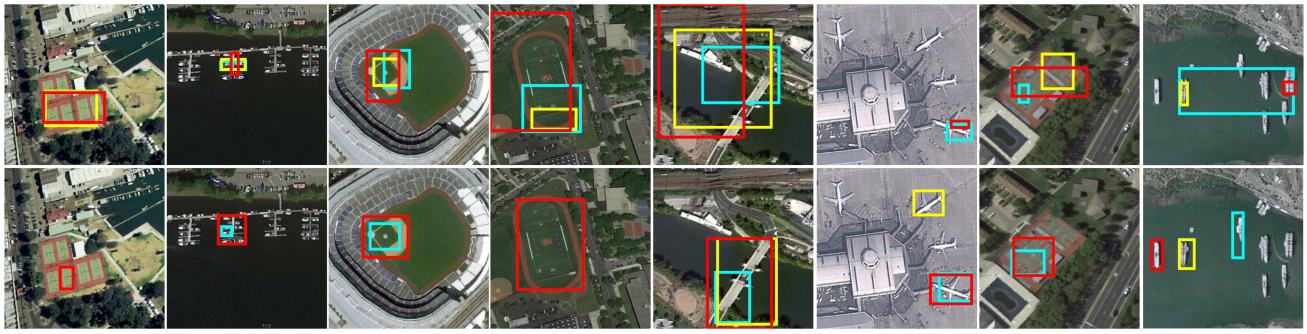


Fig. 4. Number of visualization comparisons between the top-scoring proposals selected by OICR (first rows) and our method (second rows) where cyan, yellow, and red rectangles denote the results obtained with three refinement times.

allows the model to converge better and faster or the false positive examples make the model drift. To avoid these phenomena, we will use all the proposals R that generated by the Selective Search [20] to train the network at the beginning of training. Besides, the self-pruning will be repeated n times, and then, P_{sp}^{n} are used as proposals for the remaining training process. Following this way, the proposals will be adaptively pruned as the network performance improves.

IV. Experiments

In this section, we first give detailed descriptions of our experimental settings, including data sets, evaluation metrics, and implementation details. Ablation experiments will be presented to analyze the impact of each component of our method. The complexity of our proposed method is further evaluated by running time. Finally, comparisons with the state of the arts are provided.

A. Data Sets and Evaluation Metrics

We evaluate our WSOD method on the challenging NWPU VHR-10.v2 and DIOR data sets [9]. The NWPU VHR-10.v2 consists of 1172 images (400×400 pixels) from ten object categories. This data set is composed of three groups: train, test, and valuation set. In our experiments, only the training set that contains 679 remote sensing images is employed as train data. The DIOR [9] data set contains 23 463 images and 192 472 instances of 20 object classes, which also are divided into train, test, and valuation set. For testing, two standard metrics are used to evaluate the performance of WSOD for remote sensing images. On the one hand, we evaluate our model on the testing set by measuring the mean average precision (mAP). On the other hand, our model's localization accuracy will be evaluated by calculating the correct location (CorLoc). The abovementioned two metrics treat each detection result as a positive detection when the IoU between the ground truth and bounding box is greater than 0.5, which is consistent with the PASCAL VOC criteria.

B. Implementation Details

We utilize VGG16 [48] as our backbone network, of which the last max-pooling is replaced by RoIpooling [46]. To protect the features of small-sized objects, we use dilated convolutional layers to replace the fourth max-pooling layer (pooling4) and its subsequent convolutional layers. In terms of initialization, we pretrain the backbone network on the ImageNet [49], and the newly added layers are initialized by the Gaussian distribution with a mean of 0 and the standard deviation of 0.01. During training, the network performs 30k iterations. We set the initial learning rate to 0.001 and reduce it to one-tenth of the previous rate after every 10k iterations. For each iteration, the minibatch size is set to 2 for stochastic gradient descent optimizer. The momentum and weight decay are set to 0.9 and 0.0005, respectively.

Before training the network, we use the Selective Search [20] to generate about 2000 proposals for each image. We augment the data by horizontally mirroring each image and rotating them with 90° and 180°. In addition, we use the same five scales {480, 576, 688, 864, 1200} as WSDDN [11] and use the same refinement times $K = 3$ as OICR [17]. In the DCIR, we measure the correlation coefficient between the candidate and other proposals using IoU, i.e.,$\lambda = 0.5$. At the beginning of training, we set the confidence threshold T_c for airplane, baseball diamond, basketball court, bridge, ground track field, harbor, ship, storage tank, tennis court, and vehicle as 0.5, 0.5, 0.6, 0.6, 0.8, 0.7, 0.5, 0.5, 0.6, and 0.5, respectively. Besides, the threshold growth factor k_c are set as 0.05, 0.05, 0.04, 0.04, 0.01, 0.02, 0.05, 0.05, 0.04, and 0.05, respectively. The Pth is set to 0.1. Before calculating the average precision (AP) and CorLoc, the NMS [47] is used to reduce duplicated bounding boxes with 0.3 IoU threshold.

C. Ablation Experiments

To evaluate the effectiveness of our PCIR method, the ablation experiments are constructed to analyze the effects of the key component of DCIR, such as context fusion, similarity estimation, and the proposed PPSP strategy.

1) Context Fusion: To demonstrate the power of the context fusion, we design a training strategy similar to OICR [17] that only selects the highest credible candidate to refine the next classifier. As shown in Fig. 4, our method can better mine the entire object. Benefiting from that, as shown in Table I, the performance in terms of mAP is boosted from 34.5% to 42.9%, and CorLoc is boosted from 40.0% to 61.9%, which

TABLE I
RESULTS ON THE NWPU VHR-10.v2 FOR DIFFERENT STYLES OF SIMILARITY ESTIMATION

Style of estimation	mAP(%)	CorLoc(%)
Baseline	34.5	40.0
Select highest + context fusion	42.9	61.9
Static + context fusion	49.7	65.4
Dynamic + context fusion	51.0	68.6

TABLE II
RESULTS AND SPEED ON THE NWPU VHR-10.v2 FOR DIFFERENT METHODS

Method	mAP(%)	speed(FPS)
Baseline	34.5	1.11
DCIR	51.0	0.78
DCIR+PPSP	55.0	1.04

further confirms that context fusion is effective to mine objects under weakly supervised settings.

2) Similarity Estimation: To disclose the contribution of context similarity estimation, we construct ablation experiments by using adaptive static threshold [in this way, we select the confident candidates by adaptively computing the credibility level with (8) and fixing the threshold T_c in (9)] and using dynamic adaptive growth threshold to select confident candidates. As shown in Table I, context similarity estimation can greatly improve performance. These two methods increase the mAP with 6.8% and 8.1%, respectively. Interestingly, our dynamic adaptive growth threshold is more effective for the object classes with multiple instances in an image (e.g., airplane, baseball, and ship). The main reason is that OICR [17] and the process of context fusion only select the top-scoring proposals as the next stages supervision information even though multiple instances of the same class exist in an image. However, our local-to-global context similarity estimation endeavors to mine all possible instances existing in an image. The combination of context fusion and local-to-global context similarity estimation leads to that more positive instances and more accurate supervisions are mined.

The cooperation of context fusion and similarity estimation formulates our DCIR. As shown in Table II, exploiting DCIR can bring about an accuracy improvement of 16.5% (51.0% versus 34.5%) in terms of mAP.

3) PPSP Strategy: As shown in Table II, adding the PPSP module not only improves the speed of network learning (from 0.78 to 1.04 frames/s) but also further boosts the performance (mAP from 51.0% to 55.0%). The reason is that the PPSP strategy rejects well-classified negative proposals and prevents overfitting caused by a single training sample. In addition, the training speed is accelerated due to the reduction in the number of candidates.

DCIR and PPSP formulate our PCIR, where DCIR first boosts the detection performance in mAP from 34.5% to 51.0%, and PPSP can further bring about an accuracy improvement of 3.97% and, simultaneously, improve the speed of training.

4) Settings of Pruning Times and Self-Pruning Times: In the third experiment mentioned earlier, we prove the validity

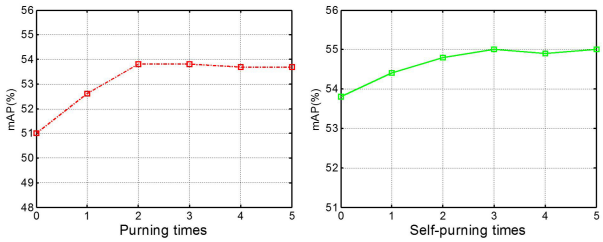


Fig. 5. Results on the NWPU VHR-10.v2 data set with (Left) different pruning times and (Right) different self-pruning times.

of the pruning process. The left of Fig. 5 shows that the pruning process can effectively improve the performance, and the best performance is achieved when two times of pruning are performed. Besides, at the beginning of each procedure, an additional initial epoch is needed. Thus, the proposal reset is operated every three epochs. However, when the pruning is executed too many times, the performance will be appropriately dropped. This is because more pruning times reduce the number of training examples and the diversity of object proposals so that the networks would be easily trapped into the local optimum.

Meanwhile, we further prove the validity of the self-pruning strategy. As shown in the right of Fig. 5, the proposal self-pruning strategy can improve the performance, but when the self-pruning is executed too many times, the results tend to be saturated (after three times of self-pruning, the improvement is small). The reason is that when the learning ability reaches a certain level, the proposals for each filtering are similar. Therefore, in order to further improve the efficiency of network learning, we only implement three times of self-pruning.

D. Running Time

The speed of our refinement algorithm is presented in Table II. Running all experiments takes about 8.5 h on ubuntu16.04, NVIDIA GTX TitanX GPU, cuDNN v5, and CUDA 8.0. As shown in Table II, compared with the baseline method, the computational efficiency dropped from 1.11 to 0.78 frames/s. According to our analysis, the additional calculations are mainly introduced by (6). For each confident candidate, we need to calculate its correlation coefficient with the remaining about 2000 proposals. This will reduce the computational efficiency from 1.11 to 0.85 frames/s but, in turn, bring about an accuracy improvement of 16.48% mAP. Furthermore, our PPSP algorithm can effectively remove redundant cluttered proposals, which can improve the computational efficiency from 0.78 to 1.04 frames/s. Although the baseline work is slightly faster than our refinement algorithm (1.11 versus 1.04 frames/s), its accuracy reduces by 16.48% compared with ours. In summary, our refinement algorithm achieves the best tradeoff between the accuracy and the speed.

E. Comparisons With State of the Arts

Here, we compare the performance of the proposed method with both weakly and fully supervised state-of-the-art methods.

TABLE III

PERFORMANCE COMPARISONS (AP AND mAP) AMONGI DIFFERENT OBJECT DETECTION METHODS ON THE NWPU VHR-10.v2 TEST SET

	Methods	Airplane	Ship	Storage tank	Baseball Diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	mAP
Fully supervised Methods	COPD [50]	0.6225	0.6937	0.6452	0.8213	0.3413	0.3525	0.8421	0.5631	0.1643	0.4428	0.5488
	Transferred CNN [7]	0.6603	0.5713	0.8501	0.8093	0.3511	0.4552	0.7937	0.6257	0.4317	0.4127	0.5961
	RICNN [30]	0.8871	0.7834	0.8633	0.8909	0.4233	0.5685	0.8772	0.6747	0.6231	0.7201	0.7311
	RCNN [51]	0.8537	0.8888	0.6278	0.1973	0.9066	0.5823	0.6795	0.7987	0.5422	0.4992	0.6576
	Fast RCNN [46]	0.9091	0.9060	0.8929	0.4732	1.0000	0.8585	0.8486	0.8822	0.8029	0.6984	0.8271
	Faster RCNN [52]	0.9090	0.8630	0.9053	0.9824	0.8972	0.6964	1.0000	0.8011	0.6149	0.7814	0.8451
	RICO [10]	0.9970	0.9080	0.9061	0.9291	0.9029	0.8013	0.9081	0.8029	0.6853	0.8714	0.8712
Weakly supervised methods	WSDDN [11]	0.3008	0.4172	0.3498	0.8890	0.1286	0.2385	**0.9943**	0.1394	0.0192	0.0360	0.3512
	OICR [17]	0.1366	0.6735	**0.5716**	0.5516	0.1364	0.3966	0.9280	0.0023	0.0184	0.0373	0.3452
	Ours	**0.9078**	**0.7881**	0.3640	**0.9080**	**0.2264**	**0.5216**	0.8851	**0.4236**	**0.1174**	**0.3549**	**0.5497**

TABLE IV

PERFORMANCE COMPARISONS (CorLoc) AMONG DIFFERENT METHODS ON THE NWPU VHR-10.v2 TRAIN SET

Methods	Airplane	Ship	Storage tank	Baseball Diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	CorLoc
WSDDN [11]	0.2232	0.3681	0.3995	0.9248	0.1796	0.2424	**0.9926**	0.1483	0.0169	0.0289	0.3524
OICR [17]	0.2941	0.8333	0.2051	0.8176	0.4085	0.3208	0.866	0.0741	0.0370	0.1444	0.4001
Ours	**1.0000**	**0.9306**	**0.6410**	**0.9932**	**0.6479**	**0.7925**	0.8969	**0.6296**	**0.1326**	**0.5222**	**0.7187**

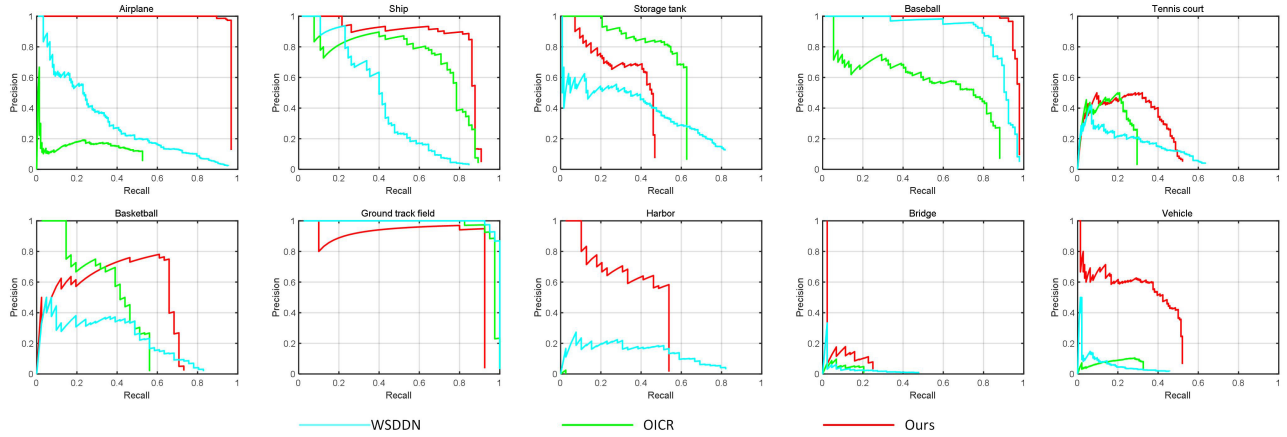


Fig. 6. PRCs of the proposed PCIR method and other WSOD approaches for each class on the NWPU VHR-10.v2 data set.

For each object class, we first provide our results on the NWPU VHR-10.v2 data set in Tables III and IV. As shown in Table III, our method achieves state-of-the-art results and even much outperforms other weakly supervised methods. The large improvement is mainly due to the following aspects.

1) The context fusion can first shift the focus of the model from the most significant parts to the entire objects so that tighter and much accurate pseudo instance-level annotations can be used to guide the refinement network.
2) All possible instances in each image are accepted through the local-to-global context fusion strategy. This method successfully copes with the characteristics of multiple objects in remote sensing images.
3) Self-pruning strategy reduces the negative influence of noise on learning and, successfully, prevents detection network fitting to simple samples.

We can see the performance measured in terms of CorLoc in Table IV. Comparing with WSDDN [11] and OICR [17], our method achieves 36% and 31% improvements, respectively. The main reason is that remote sensing images always contain multiple instances. Only selecting the top-scoring proposal as a positive example leads to a lot of information loss. Due to the mining of as many instances as possible in the images, we can make great progress.

We also compare our results with fully supervised object detection methods designed for remote sensing images, such as the collection of part detector(COPD) [50], the transferred CNN model from AlexNet [7], the RICNN [30], RCNN [51], Fast-RCNN [46], and Faster-RCNN [52]. As shown in Table II, our method further narrows the gap between the weakly and full supervised methods. Especially, for the object classes of airplane, baseball, and ground track field, our method achieves comparable or even superior results compared with fully supervised learning methods.

The precision–recall curves (PRCs) of three different WSOD methods, as shown in Fig. 6, further demonstrate the superiority of PCIR over the existing state-of-the-art methods.

Tables V and VI present the comparisons of our method with both weakly and fully supervised state-of-the-art methods in terms of mAP and CorLoc on the DIOR data set,

Fig. 7. Example results on the NWPU VHR-10.v2 test split (54.97% mAP) for each class. The first two rows indicate success cases and different colors rectangle indicates different classes. The third rows correspond to a few missed objects and false positives.

TABLE V

PERFORMANCE COMPARISONS (AP AND mAP) AMONG DIFFERENT METHODS ON THE DIOR TEST SET

Methods	Fully supervised methods		Weakly supervised methods		
	Fast RCNN [46]	Faster RCNN [52]	WSDDN [11]	OICR [17]	Ours
Airplane	0.4417	0.5028	0.0906	0.0870	**0.3037**
Airport	0.6679	0.6260	**0.3968**	0.2826	0.3606
Baseball field	0.6696	0.6604	0.3781	0.4405	**0.5422**
Basketball court	0.6049	0.8088	0.2016	0.1822	**0.2660**
Bridge	0.1556	0.2880	0.0025	0.0130	**0.0909**
Chimney	0.7228	0.6817	0.1218	0.2015	**0.5859**
Dam	0.5195	0.4726	**0.0057**	0.0009	0.0022
Expressway service area	0.6587	0.5851	0.0065	0.0065	**0.0965**
Expressway toll station	0.4476	0.4806	0.1188	0.2989	**0.3618**
Golf field	0.7211	0.6044	0.0490	0.1380	**0.3259**
Ground track field	0.6293	0.6700	0.4235	0.5739	**0.5851**
Harbor	0.4618	0.4386	0.0466	**0.1066**	0.0860
Overpass	0.3803	0.4687	0.0106	0.1106	**0.2163**
Ship	0.3213	0.5848	0.0070	0.0909	**0.1209**
Stadium	0.7098	0.5237	0.6303	0.5929	**0.6428**
Storage tank	0.3504	0.4235	0.0395	0.0710	**0.0909**
Tennis court	0.5827	0.7952	0.0606	0.0068	**0.1362**
Train station	0.3791	0.4802	**0.0051**	0.0014	0.0030
Vehicle	0.1920	0.3477	0.0455	0.0909	**0.0909**
Windmill	0.3810	0.6544	0.0114	0.0041	**0.0752**
mAP	0.4998	0.5548	0.1326	0.1650	**0.2492**

TABLE VI

PERFORMANCE COMPARISONS (CorLoc) AMONG DIFFERENT WSOD METHODS ON THE DIOR TRAIN SET

Methods	CorLoc
WSDDN [11]	0.3244
OICR [17]	0.3477
Ours	**0.4612**

respectively. It is obvious that PCIR significantly outperforms the other two weakly supervised methods by 11.66% mAP and 8.42% mAP. Our method narrows the gap between the weakly supervised and fully supervised methods, especially for the object classes of "chimney" and "stadium." For the classes of airplane, baseball filed, chimney, golf filed, and overpass, the detection accuracies are significantly improved by larger than 10% mAP. Meanwhile, compared with the state-of-the-art WSOD methods, our PCIR brings about the improvements of 13.68% and 11.35% CorLoc.

The abovementioned comparison clearly demonstrates the effectiveness of our method. Although our approach has made great progress, our results for some classes are still not satisfactory, such as bridges. There are two main reasons: 1) the appearances of roads are particularly similar to bridges and 2) lacking instance-level supervision and the coexistence of bridges and rivers lead to the detection network misunderstanding the rivers as bridges. Similarly, the coexistence of reservoirs and dams (windmills and their shadows) leads to the detection network misunderstanding reservoirs as dams (windmills' shadows as windmills). Compared with the results of fully supervised methods, the performance of the object classes of the airport, expressway service area, expressway toll station, golf filed, storage tank, tennis court, and vehicle needs to be further improved. The main reasons for the large gap in these classes are as follows: 1) the object is only a very small proportion relative to the background, especially the vehicles and 2) these categories often appear in adjacent locations, causing the model to mistake multiple objects for one. Thus, an ideal solution is still needed to improve the performance of detection.

F. Qualitative Results

Some detection results on the NWPU VHR-10.v2 and DIOR data sets are presented in Figs. 7 and 8, respectively. As can be seen, our method can give accurate and tight bounding boxes to each instance appearing in the images. However, for some classes (the last rows in Figs. 7 and 8), such as bridge, dam, and windmill, our method sometimes fails by misunderstanding rivers as the bridges, reservoirs as dams, and windmills' shadows as windmills or only detects the discriminative parts of objects. If we incorporate the attention mechanism into our framework, the performance of detection will be further improved.

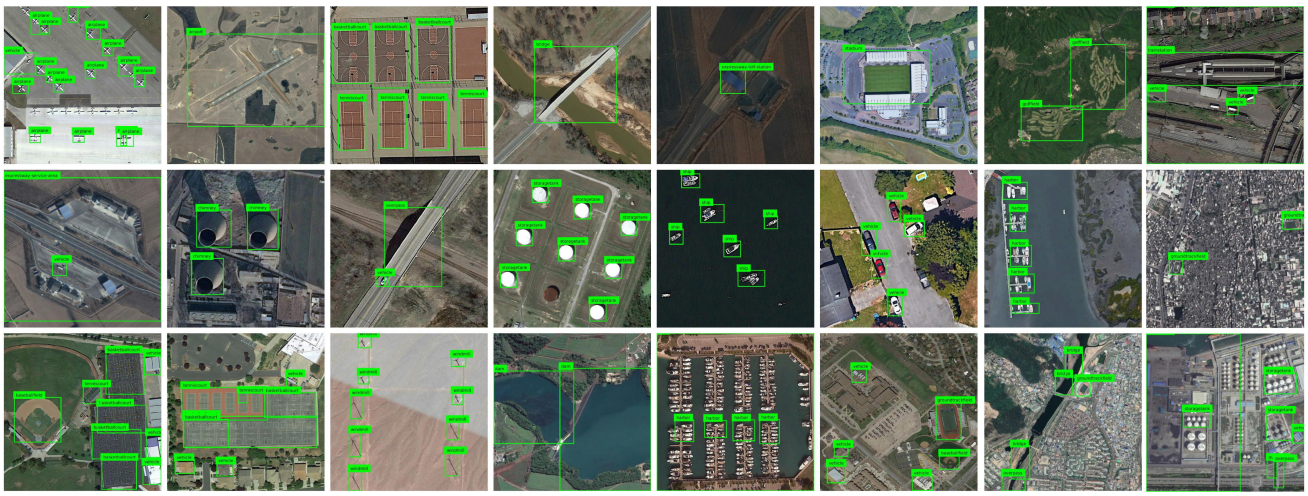


Fig. 8. Example results on the DIOR test split (24.92% mAP). The first two rows indicate success cases. The third rows correspond to a few missed objects and false positives.

V. Conclusion

We presented an end-to-end online PCIR approach to handle the WSOD problem in remote sensing images. A DCIR strategy is designed to divert the focus of detection network from local distinct part to the object and further to other potential instances by leveraging both local and global context information. Benefiting from DCIR, a PPSP algorithm is further developed to mitigate the influence of complex background by dynamically rejecting the negative training proposals. Extensive experiments clearly demonstrated that the proposed method can significantly boost object detection accuracy compared with the state of the arts.

References

[1] S. Tuermer, F. Kurz, P. Reinartz, and U. Stilla, "Airborne vehicle detection in dense urban areas using HoG features and disparity maps," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 6, pp. 2327–2337, Dec. 2013.

[2] P. Zhong and R. Wang, "A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3978–3988, Dec. 2007.

[3] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, Oct. 2018.

[4] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3196–3209, Jul. 2017.

[5] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[6] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul. 2019.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[9] 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.

[10] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

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

[12] 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.

[13] 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.

[14] 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.* Springer, 2016, pp. 350–365.

[15] 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.

[16] 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.

[17] 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. 3059–3067.

[18] 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.

[19] 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. 4294–4302.

[20] 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.

[21] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 391–405.

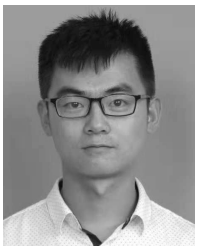
[22] X. Zhang, J. Feng, H. Xiong, and Q. Tian, "Zigzag learning for weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4262–4270.

[23] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 524–531.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[25] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[26] J. Han *et al.*, "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogramm. Remote Sens.*, vol. 89, pp. 37–48, Mar. 2014.
[27] G. Cheng, J. Han, L. Guo, and T. Liu, "Learning coarse-to-fine sparselets for efficient object detection and scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1173–1181.
[28] L. Zhang, L. Zhang, D. Tao, and X. Huang, "Sparse transfer manifold embedding for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1030–1043, Feb. 2014.
[29] Y. Zhang, B. Du, and L. Zhang, "A sparse representation-based binary hypothesis model for target detection in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1346–1354, Mar. 2015.
[30] 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.
[31] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," *Sensors*, vol. 17, no. 2, p. 336, 2017.
[32] Y. Yang, Y. Zhuang, F. Bi, H. Shi, and Y. Xie, "M-FCN: Effective fully convolutional network-based airplane detection framework," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1293–1297, Aug. 2017.
[33] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
[34] 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.
[35] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.
[36] P. Zhou, G. Cheng, Z. Liu, S. Bu, and X. Hu, "Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping," *Multidimensional Syst. Signal Process.*, vol. 27, no. 4, pp. 925–944, Oct. 2016.
[37] X. Zeng *et al.*, "Crafting GBD-net for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2109–2123, Sep. 2018.
[38] R. Yu, X. Chen, V. Morariu, and L. Davis, "The role of context selection in object detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–13.
[39] W. Ouyang, K. Wang, X. Zhu, and X. Wang, "Learning chained deep features and classifiers for cascade in object detection," 2017, *arXiv:1702.07054*. [Online]. Available: https://arxiv.org/abs/1702.07054
[40] S. Zagoruyko *et al.*, "A MultiPath network for object detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–14.
[41] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1134–1142.
[42] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.
[43] Z. Chen, S. Huang, and D. Tao, "Context refinement for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 71–86.
[44] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.
[45] Y. Wei *et al.*, "Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 434–450.
[46] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[47] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6469–6477.
[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[49] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[50] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.
[51] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[52] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.



Xiaoxu Feng received the B.E. degree from Inner Mongolia University, Hohhot, China, in 2017. He is pursuing the Ph.D. degree with Northwestern Polytechnical University, Xi'an, China.

His research interests include computer vision and remote sensing image processing, especially on object detection and scene classification.

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, Singapore, The Chinese University of Hong Kong, Hong Kong, Dublin City University, Dublin, Ireland, and the University of Dundee, Dundee, U.K., from 2003 to 2010. He is a Professor with Northwestern Polytechnical University. His research interests include computer vision and brain-imaging analysis.

Dr. Han is also an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON MULTIMEDIA.

Xiwen Yao received the B.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2010 and 2016, respectively.

He is a Research Assistant with Northwestern Polytechnical University. His research interests include computer vision and remote sensing image processing, especially on fine-grained image classification and object detection.

Gong Cheng (Member, IEEE) 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 a Professor with Northwestern Polytechnical University. His main research interests include computer vision and pattern recognition.