Self-Guided Proposal Generation for Weakly Supervised Object Detection

Gong Cheng, *Member, IEEE*, Xuan Xie, Weining Chen, Xiaoxu Feng, Xiwen Yao, *Member, IEEE*, and Junwei Han, *Fellow, IEEE*

***Abstract*— Weakly supervised object detection (WSOD) in remote sensing images remains a challenging task when learning object detectors with only image-level labels. As we know, object proposal generation plays a crucial role in WSOD. At present, the proposal generation of most existing WSOD methods mainly relies on heuristic strategies such as selective search and Edge Boxes. However, the proposals obtained by the above methods cannot well cover the entire objects, severely hindering the performance of WSOD. To address this issue, this article proposes a Self-guided Proposal Generation approach, termed SPG. It can be easily implemented with most WSOD methods in a unified framework. To this end, we first introduce a confidence propagation approach to obtain the objectness confidence map for each image, which, on the one hand, highlights informative object locations and, on the other hand, aggregates discriminative feature representation by combining the objectness confidence map with the deep features. Then, the proposal generation is implemented by mining informative regions as proposals on the objectness confidence map. Extensive evaluations on two challenging datasets demonstrate that our SPG significantly improves the baseline methods, online instance classifier refinement (OICR) and min-entropy latent model (MELM), by large margins (for OICR: 15.86% mAP and 12.89% CorLoc gains on the NWPU VHR-10.v2 dataset and 3.65% mAP and 4.87% CorLoc gains on the DIOR dataset; for MELM: 20.51% mAP and 23.54% CorLoc gains on the NWPU VHR-10.v2 dataset and 7.11% mAP and 4.96% CorLoc gains on the DIOR dataset) and achieves the state-of-the-art results compared with existing methods.**

***Index Terms*— Proposal generation, remote sensing images (RSIs), weakly supervised object detection (WSOD).**

I. Introduction

OBJECT detection in remote sensing images (RSIs) is a fundamental yet challenging task, which aims to locate and classify object instances in aerial or satellite images. With the advances in convolutional neural network (CNN) [1], considerable success has been achieved in object detection in the field of remote sensing [2]–[13] due to the availability of large scale datasets with accurate instance-level annotations. However, annotating objects with bounding boxes is both prohibitively expensive and time-consuming. By contrast, image-level annotations are generally much cheaper to acquire, which drives many researchers to explore object detection methods under the weakly supervised setting [14]–[23]. This is also known as weakly supervised object detection (WSOD). It aims to learn object detectors with only the image-level annotations indicating whether an object exists in an image.

Currently, the most popular pipeline for WSOD follows a two-phase learning procedure: proposal generation and proposal classification. The proposal generation phase divides images into a series of candidate boxes as region proposals that may contain objects. The proposal classification phase iteratively selects confident positive proposals as pseudo instance-level labels and trains object detectors to classify each proposal as an object or background under multiple instance learning (MIL) constraints [24]. One of the pioneering works by Bilen and Vedaldi [25] first developed an end-to-end weakly supervised deep detection network (WSDDN) for WSOD, in which the final image classification score is a weighted sum of the proposal scores, that is, each proposal contributes a certain percentage to the final image classification. Based on the WSDDN, a number of follow-up works [26]–[42] further boost the performance of WSOD through different strategies, e.g., leveraging spatial relations [26], [28], [35], better optimization [30], [34], [36], [37], and multitasking with weakly supervised segmentation [38]–[40].

Nevertheless, the proposal generation of the abovementioned studies usually adopts heuristic strategies for WSOD, such as selective search (SS) [43] and Edge Boxes (EB) [44]. However, the proposals obtained by the above methods cannot well cover the entire objects, resulting in that the Intersection over Union (IoU) values between the proposals and the ground-truth boxes are small, as shown in Fig. 1(a). This inability of generating high-quality proposals severely affects the precise object localization for WSOD. Thus, the problem of how to generate high-quality proposals for WSOD remains open. The famous work of the region proposal network (RPN) [45] has proved that the CNN-based region proposal generation approach is an indispensable component in advanced two-stage fully supervised object detectors. However, to ensure high performance, RPN requires instance-level annotations to train the network, leading to that this approach

Manuscript received February 5, 2022; revised May 4, 2022; accepted June 3, 2022. Date of publication June 8, 2022; date of current version June 23, 2022. This work was supported in part by the National Science Foundation of China under Grant 62136007, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021B1515020072, and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Gong Cheng.)*

Gong Cheng and Xuan Xie are with the Research and Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China, and also with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: chenggong1119@gmail.com).

Weining Chen, Xiaoxu Feng, Xiwen Yao, and Junwei Han are with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China.

Digital Object Identifier 10.1109/TGRS.2022.3181466

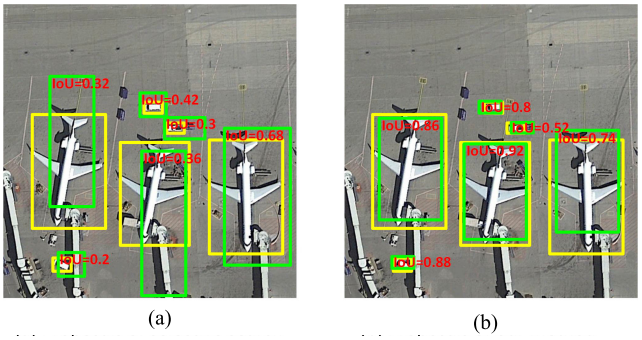


Fig. 1. IoU between the top-scoring proposals (green boxes) and the ground-truth boxes (yellow boxes). (a) Proposals of SS cannot fully cover the ground-truth boxes. (b) Proposals of our method have higher quality than those obtained by SS [43].

cannot be directly applied to WSOD because only image-level annotations are available during training. In addition, Tang *et al.* [29] proposed a two-stage RPN for proposal generation for WSOD. It first evaluates the objectness scores of sliding window boxes to generate coarse proposals by following the EB [44] and then refines the proposals using a region-based CNN classifier. Nevertheless, this method is not efficient because the number of sliding window boxes is usually very large (hundreds of thousands or even millions).

The above analyses inspire us to design CNN feature-based proposal generation method to generate high-quality proposals for boosting the detection performance of WSOD. To this end, we put forward a Self-guided Proposal Generation approach, termed SPG, for WSOD. Compared to handcrafted features used in heuristic proposal generation strategies, such as SS [43] and EB [44], we argue that high-level CNN features, which contain rich semantic information about the objects, are more suitable for object proposal generation. More specifically, in order to explore more semantic information and discover more reliable object regions, a high-level features-based confidence propagation method is first designed to obtain the objectness confidence map, which highlights informative object locations, by random walk strategy. Note that the random walk can use local information to learn latent representations in unsupervised or weakly supervised feature learning [46]–[49]. The random walk fully leverages the contextual information to propagate objectness confidence by calculating the similarity between the locations in the deep feature maps. Then, we mine informative regions as proposals, which contain potential objects, on the objectness confidence map. As shown in Fig. 1(b), the IoU between the proposals of our SPG method and the ground-truth boxes is larger than SS [43], thus better covering the whole objects. This suggests that our SPG method could generate higher quality proposals than that of SS [43]. In addition, to further aggregate more information for improving the feature representation capability, we consider the above-obtained objectness confidence map as a spatial attention clue. Thus, the enhanced feature maps can be obtained by the Hadamard product of the objectness confidence map with high-level features, which can highlight informative object regions and suppress useless background information.

We elaborately conduct a large number of experiments on the challenging NWPU VHR-10.v2 and DIOR datasets to verify the effectiveness of our SPG method. To sum up, our contributions are threefold as follows.

1) We rethink the proposal generation of WSOD. Instead of generating proposals through SS [43] and EB [44], we design a novel SPG method with the confidence propagation strategy based on deep CNN features. The proposal generation and WSOD network can be integrated into a unified network for joint optimization.
2) Our proposed SPG can be easily inserted into many popular WSOD frameworks. We implement our SPG method with online instance classifier refinement (OICR) [26] and min-entropy latent model (MELM) [27] and significantly improve the overall performance of both two baseline methods.
3) Experiments show that our proposed SPG method obtains better detection results over previous state-of-the-art methods on the challenging NWPU VHR-10.v2 and DIOR benchmarks for WSOD.

II. Related Work

A. Object Detection in Remote Sensing Images

During the last decade, with the rapid development of deep learning, several CNN-based fully supervised object detection approaches [2]–[13] have achieved remarkable results in RSIs. For instance, Cheng *et al.* [3] proposed a novel and effective approach to learn a rotation-invariant CNN (RICNN) model for gaining the huge performance of object detection, which tackles the variations of object orientation in RSIs. Then, Cheng and Han [5] employed a simple but convincing method to train rotation-invariant and Fisher discriminative CNN models to further achieve better detection results based on the existing state-of-the-art object detection systems. Li *et al.* [7] designed a hyper RPN (HRPN) and a cascade of boosted classifiers for accurately detecting vehicles in RSIs. Cheng *et al.* [8] proposed a multi-scale object proposal network (MS-OPN) and an accurate object detection network (AODN) for simultaneously detecting multiclass objects in RSIs with large scales variability. Wu *et al.* [9] made a comprehensive review of the recent deep learning-based object detection progress in both the computer vision and earth observation communities, and proposed a publicly available benchmark for object detection in RSIs. Although the fully supervised learning paradigm has achieved great success, its training relies on instance-level annotations, e.g., tight bounding boxes, to support satisfactory detection.

Therefore, weakly supervised learning using only image-level annotations has attracted a great deal of attention. Recently, many efforts have been made to develop various approaches for learning object detectors with weak supervision in RSIs. For instance, Han *et al.* [14] first attempted to address the WSOD problem in RSIs, which adopts deep Boltzmann machines (DBMs) to learn high-level features and utilizes a new weakly supervised learning framework based on Bayesian principles to detect objects from optical RSIs. Zhou *et al.* [15] proposed a weakly supervised approach that

transfers a deep model to extract high-level features from RSIs and integrates a negative bootstrapping scheme into the detector training process. Zhang *et al.* [16] developed a novel framework that consists of training set initialization and target detector learning to efficiently detect targets from RSIs. Li *et al.* [21] proposed a weakly supervised deep learning (WSDL) method for multiclass geospatial object detection using scene-level tags only, which exploits both the separate scene category information and mutual cues between scene pairs to sufficiently train deep networks for pursuing the superior object detection performance. Besides, Li *et al.* [22] also proposed a WSDL-based cloud detection (WDCD) method using block-level labels indicating only the presence or the absence of cloud in one RS image block. However, the feature representation of the above approaches is limited, which leads to a large performance gap with target detection in RSIs under fully supervised learning. Then, Yao *et al.* [17] proposed a method of dynamic curriculum learning for automatic WSOD from high-spatial-resolution RSIs. Feng *et al.* [18] designed a progressive contextual instance refinement method for WSOD in RSIs, in which a dual-contextual instance refinement (DCIR) module and a progressive proposal self-pruning (PPSP) strategy are developed to boost the detection accuracy. Besides, Feng *et al.* [19] introduced a triple context-aware network (TCA-Net) to capture discriminative and supplementary cues for WSOD. It includes a global context-aware enhancement (GACE) module and a dual-local context residual (DLCR) module to tackle the challenges of local optimization and missing adjacent instances. Wang *et al.* [20] proposed a dynamic pseudolabel generation framework by using the localization information to generate pseudolabels for each proposal. Although the abovementioned studies have achieved promising results, they still utilize SS [43] to obtain object proposals, thus severely limiting the performance of WSOD.

B. Region Proposal Generation

It is obvious that the performance of object detection heavily depends on the quality of the object proposals. The higher the quality of proposals, the more the accurate the object detection. There are several works focusing on region proposal generation [43]–[45], where SS [43] and EB [44] are two most commonly used proposal generation methods for WSOD. Specifically, SS generates proposals based on a superpixel merging method. It carries out a hierarchical grouping algorithm on the basis of graph-based segmentation. Different from the SS algorithm based on segmentation and region similarity, the EB method divides image regions by mining the edge information of the images and evaluating the objectness scores of sliding window boxes. As we have analyzed before and illustrated in Fig. 1, most of the proposals generated by these two methods cannot well fit the ground-truth bounding boxes of objects, which results in that the learned object detectors may not well localize objects. RPN regresses object locations based on deep convolutional features [45], which has obtained state-of-the-art proposal performance for recent fully supervised object detectors. However, to ensure high performance, RPN [45] still requires bounding box annotations, which limits its applicability to WSOD. Unlike the abovementioned methods, e.g., SS [43] and EB [44], which use redundant proposals generated with handcrafted features to hypothesize objects' locations, our SPG method spotlights potential object locations via performing objectness confidence propagation over the deep feature maps. It can be plugged into any standard WSOD architecture to introduce high-quality object proposals, significantly boosting WSOD performance.

C. Random Walk

A random walk is known as a random process. It describes a path that consists of a succession of random steps in the mathematical space [46]–[49]. The random walk can be used to analyze the randomness of objects and calculate the correlation among objects in an image. It has increasingly been popular in semi-supervised and weakly supervised learning in computer vision. In this article, we generate object proposals by confidence propagation with a random walk strategy based on deep CNN features. In brief, a random walk is adopted to accumulate objectness confidence between adjacent object locations by fully considering the semantic relevance of neighboring regions.

III. Proposed Method

A. Overview

As mentioned above, this article mainly focuses on the CNN-based proposal generation method to obtain high-quality proposals, further advancing the performance of WSOD. To fulfill this purpose, we implement our method based on the OICR [26] and MELM [27] frameworks by introducing two important modules called confidence propagation and proposal generation. The architecture of our method is shown in Fig. 2. First, we introduce the confidence propagation, which aims to obtain an objectness confidence map for each given image based on the deep feature maps with a random walk strategy (see Section III-B). Then, we describe the proposal generation in detail, which aims to generate a series of object proposals on the objectness map (see Section III-C). Finally, two basic WSOD frameworks are introduced briefly, which also serve as the baseline methods (see Section III-D).

B. Confidence Propagation

Intuitively, the pixels within one object share similar features. Given image-level annotations, most previous methods modeled image-level loss as the accumulated scores over regions and performed detection based on the region scores. Nevertheless, the image-level loss is defined by classification loss, which makes the models prone to being trapped in local minimums. Consequently, it is relatively easy to find the most discriminative object areas but very difficult to find most object instances. This weakness usually results in the detection of partial bounding boxes, which, in fact, has been well verified in most WSOD methods. Motivated by this, a natural solution to this problem is propagating information of most discriminative object areas to their adjacent regions, thus capturing more

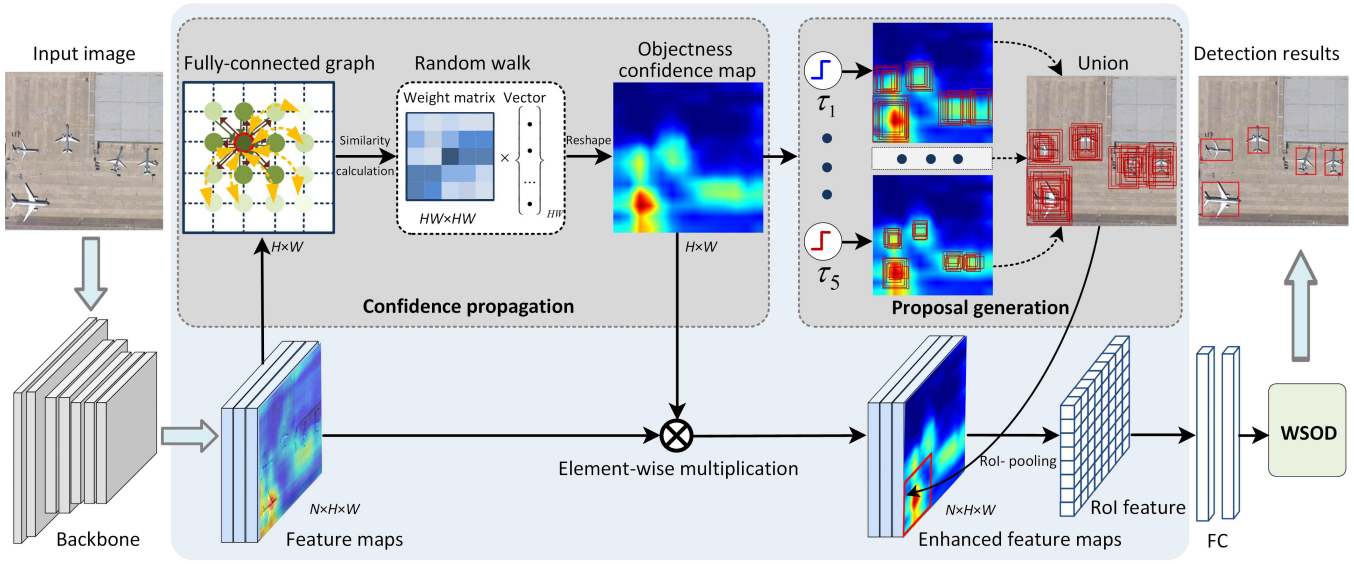


Fig. 2. Architecture of our SPG for WSOD. To implement SPG, we first adopt the confidence propagation module to generate an objectness confidence map for each image and then mine informative regions as proposals on the objectness confidence map under different thresholds. Finally, the proposals obtained by SPG are input to the WSOD network for object detection.

objects accordingly. To this end, we introduce a confidence propagation module to capture as many object regions as possible via the clue of objectness confidence. Furthermore, the objectness confidence can also serve as spatial attention to help us enhance relevant features and suppress irrelevant features during the inference process.

Let $\mathbf{F} \in \mathbb{R}^{N\times H\times W}$ represent the feature maps output by the convolutional layer of an input image, where N represents the number of channels, and $H \times W$ is the size of the feature map. Each spatial location (i, j) on $\mathbf{F}$ can be represented by a feature vector $\boldsymbol{f}_{i,j} \in \mathbb{R}^{N\times 1}$. Here, we will expound on how to propagate the confidence between every two locations on the feature maps. Based on the theoretical hypothesis: 1) the object areas of the same category have similar feature representations and 2) the adjacent object regions have a certain semantic correlation, the feature similarity and spatial distance can be combined to describe the objectness confidence.

Thus, the objectness confidence is defined as

$$\mathbf{A}_{(i_1,j_1),(i_2,j_2)} \triangleq \left\|\boldsymbol{f}_{(i_1,j_1)} - \boldsymbol{f}_{(i_2,j_2)}\right\|_2 \cdot \exp\left(-\frac{(i_1-i_2)^2+(j_1-j_2)^2}{2\sigma}\right) \tag{1}$$

where $\|\cdot\|_2$ denotes the L2-norm, $\boldsymbol{f}_{(i_1,j_1)}$ and $\boldsymbol{f}_{(i_2,j_2)}$ represent the feature vectors of the locations (i_1, j_1) and (i_2, j_2), respectively, and σ is empirically set as $0.1H$ in the experiments (H is equal to W in this article). Then, the weight matrix $\mathbf{A}' \in \mathbb{R}^{HW\times HW}$ is normalized as follows:

$$\mathbf{A}' = \frac{\mathbf{A}}{\sum \mathbf{A}} \tag{2}$$

where $\sum \mathbf{A}$ represents the sum of all elements in the matrix $\mathbf{A}$. The weight matrix $\mathbf{A}'$ represents the feature similarity and semantic correlation between two locations, which is utilized to generate the objectness confidence map. In this article, we select a random walk strategy to propagate the objectness confidence. Specifically, each image is modeled as a graph where each location corresponds to a node. In this way, the relationship of each two neighboring locations (represented by nodes) can be modeled by the edges of the graph. The weights of edges are reflected by the objectness confidence between the locations. The random walk algorithm iteratively accumulates the objectness confidences at the locations that have high similarity with their surroundings.

Compared with other graph-propagation approaches, the random walk algorithm can incorporate a great deal of contextual information. Different from the traditional random walk algorithm, in our experimental settings, we first select the location corresponding to the highest response value as the initial node, and then, the weights of the edges between the initial node and other nodes are calculated with (1) and (2). Due to the wide range of RSIs, the targets usually occupy a relatively small area. We found that it is not necessary to take a large step. Taking into account the efficiency and accuracy of the algorithm, we take 1 as the step number. According to the above random walk algorithm, each node receives the confidence from the inbound directed edges and then diffuses it along the outbound directed sides. On the whole, the objectness confidences are used in the random walk as the weights of edges so that the random walk propagates confidence to nearby areas of the same semantic entity, which improves the accuracy of object localization significantly and allows us to generate high-quality object proposals consequently.

In the process of confidence propagation, the 2D objectness confidence map $\mathbf{G}$ is first reshaped to a vector with $H \times W$ elements initialized with the value of $1/HW$. $\mathbf{G}$ is then updated by multiplying the weight matrix $\mathbf{A}'$.

$$\mathbf{G} \leftarrow \mathbf{A}' \times \mathbf{G} \tag{3}$$

where the operator "$\times$" is the matrix multiplication. The above process conforms to the eigenvector centrality theory [50]. The basic idea of eigenvector centrality is that the centrality

of a node is a function of the centrality of adjacent nodes. In other words, the importance of a node depends not only on the number of its neighbor nodes but also on the importance of its neighbor nodes. Since the weight matrix $\mathbf{A}'$ is based on the deep feature map $\mathbf{F}$, and $\mathbf{F}$ is based on the convolution kernel $\mathbf{k}$ of the convolution layers, the dependence relationship can be described as follows:

$$\mathbf{G} \leftarrow \mathbf{A}'(\mathbf{F}(\mathbf{k})) \times \mathbf{G}. \tag{4}$$

In (4), experimentally, the propagation process reaches stable in about 12 iterations. Finally, $\mathbf{G}$ is reshaped from a vector to a 2D objectness confidence map $\mathbf{G} \in \mathbb{R}^{H \times W}$.

Through the above operations, the objectness confidence map indicates the possible object regions based on the high-level CNN features. In order to enhance the representational power of the network, we focus on enhancing the spatial information of convolutional features to selectively emphasize informative features and suppress unnecessary ones in a computationally efficient manner. In short, the objectness confidence map $\mathbf{G}$ is multiplied by the feature map $\mathbf{F}$ output by the last feature layer to obtain enhanced feature maps $\mathbf{M}$

$$\mathbf{M} = \mathbf{F} \circ \mathbf{G} \tag{5}$$

where $\circ$ indicates elementwise multiplication. As a result, the enhanced feature maps are shown to be more effective in highlighting informative object regions.

Through the objectness confidence propagation by random walk and the objectness confidence map generation, the confidence propagation module spotlights potential object locations on the high-level CNN features and, thus, can be used to generate object proposal, as described in the following, for WSOD in a self-guided manner.

C. Proposal Generation

In order to obtain candidate object regions, we design a simple yet effective approach to mine the bounding boxes of potential objects on the objectness confidence map. To be specific, we first set five segmentation thresholds τ_1, τ_2, τ_3, τ_4, and τ_5 that are equally distributed between the maximum gray value and the average gray value of overall pixels on the objectness confidence map. Then, the maximum connected area (MCA) algorithm is used to obtain the minimum enclosing rectangles based on the set thresholds. After the above operation, we can generate a series of object proposals for each segmentation threshold. We aggregate all proposals obtained under different thresholds into a union. Finally, these proposals are fed into a Region-of-Interest (RoI) pooling layer and two fully connected (FC) layers to obtain proposal feature vectors for object detection based on two basic WSOD models, which will be described in the following.

D. Basic WSOD Models

Recently, many popular WSOD methods have significantly improved the performance of object detection. To verify the effectiveness of our proposed SPG method, we implement it with two basic WSOD frameworks, named OICR [26] and MELM [27].

Given an input image $\mathbf{x}$ with the image-level label $\mathbf{y} = [y_1, \ldots, y_C]$, we can obtain a list of candidate object proposals $R = \{\mathrm{R}_1, \ldots, \mathrm{R}_{|R|}\}$ by our method, where $y_c = 1$ or 0 indicates the image with or without object class c, C is the number of object classes, and $|R|$ denotes the number of proposals. To address the discriminative part domination issue of WSOD, OICR [26] introduces multiple instance classifier refinement branches to refine the multiple instance detection network (MIDN). In the MIDN module, the above-obtained feature vectors of object proposals are first branched into two parallel classification and detection streams to generate two matrices: $\mathbf{x}^{\mathrm{cls}}$ and $\mathbf{x}^{\mathrm{det}}$. Then, the two matrices are normalized by the softmax layers $\sigma(\cdot)$ along the category direction and proposal direction to produce the classification scores $\sigma(\mathbf{x}^{\mathrm{cls}})$ and detection scores $\sigma(\mathbf{x}^{\mathrm{det}})$ of each proposal. The proposal scores are computed by the elementwise product $\mathbf{x}^R = \sigma(\mathbf{x}^{\mathrm{cls}}) \odot \sigma(\mathbf{x}^{\mathrm{det}})$. Finally, the image-level classification score for the c-th class is computed as $p_c = \sum_{r=1}^{|R|} x_{\mathrm{cr}}^R$ by the sum over all proposals, where x_{cr}^R is the proposal score of the r-th proposal for the c-th class. Thus, we can train the basic instance classifier by standard multiclass cross-entropy loss, as shown in (6). The instance classifier refinement modules focus on progressively refining object detectors. Each refinement stage is supervised by the instance-level pseudo labels selected from top-scoring proposals in previous stage (the first refinement stage is based on MIDN). Using the proposal-level supervision information, for the n-th refinement, we can train the refined classifier based on the loss function in (7). After obtaining supervision and loss for training refined classifiers, we can get the loss of our overall network by combining (6) and (7) as (8). N is the refinement times. Through optimizing this loss function, the object detectors can detect larger parts of objects gradually. More details can be found in [26].

$$L_1 = -\sum_{c=1}^{C} (y_c \log p_c + (1 - y_c) \log(1 - p_c)) \tag{6}$$

$$L_2^n = -\frac{1}{|R|} \sum_{r=1}^{|R|} \sum_{c=1}^{C+1} y_{\mathrm{cr}}^n \log x_{\mathrm{cr}}^{Rn} \tag{7}$$

$$L = L_1 + \sum_{n=1}^{N} L_2^n. \tag{8}$$

For the purpose of reducing the variance of learned instances and alleviating the ambiguity of weakly supervised object detectors, MELM [27] was proposed recently. It is decomposed into three components, including proposal clique partition, object clique discovery, and object localization. According to spatial and category relations, the object proposals are divided to construct the object cliques for reducing the proposal redundancy. To minimize localization randomness, a global min-entropy model that reflects the spatial and category distributions of object cliques is designed as $E_d(h, \Theta)$, which can be described as follows:

$$\begin{aligned} E_d(h, \Theta) &= -\log \sum_c w_{H_c} p_{H_c} \\ &= -\log \sum_c w_{H_c} \sum_{h \in H_c} p(y, h; \Theta) \end{aligned} \tag{9}$$

where c is the number of cliques, $y \in \{0, 1\}$ denotes the class label indicating whether the input image contains an object or not, $w_{H_c} = 1/|H_c| \sum_{h \in H_c} (p(y, h; \Theta)/\sum_y p(y, h; \Theta))$ measures the probability distribution of objects to all classes in a spatial clique H_c, Θ represents network parameters, and $p(y, h; \Theta)$ is the joint probability of class y and latent variable (object proposal) h. To accurately localize the objects in the discovered cliques, a local MELM is designed as $E_l(h, \Theta)$, which can be described as follows:

$$E_l(h, \Theta) = -\log \max_{h \in H_c^*} w_h \cdot p(y, h; \Theta) \tag{10}$$

where H_c^* represents the clique with the highest average object confidence and $w_h = p(y, h; \Theta)/\sum_y p(y, h; \Theta)$ measures the distribution of object confidences to all image classes y. Thus, the whole MELM is defined as follows:

$$\{h^*, \Theta^*\} = \arg\min_{h, \Theta} E_d(h, \Theta) + E_l(h, \Theta). \tag{11}$$

More explanations can be found in [27].

IV. Experiments

In this section, the datasets and evaluation metrics are first described in detail, followed by the introduction of implementation details. Next, the ablation experiments are designed to analyze the contribution of our proposed SPG method. Finally, a series of quantitative and qualitative comparisons are made with existing state-of-the-art works.

A. Datasets and Evaluation Metrics

We comprehensively evaluate our method on two challenging datasets: NWPU VHR-10.v2 [6] and DIOR [9]. Specifically, these two datasets are the widely used benchmarks for WSOD in RSIs. NWPU VHR-10.v2 contains a total of 1172 images with the size of 400×400 pixels, covered by ten object categories, namely, airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. Both the training set (679 images) and the validation set (200 images) are used for training, and the remaining (293 images) are used for testing. DIOR is a large-scale dataset with the image size of 800×800 pixels, covered by the following 20 object classes: airplane (APL), airport (APO), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), dam (DAM), expressway service area (ESA), expressway toll station (ETS), golf field (GF), ground track field (GTF), harbor (HA), overpass (OP), ship (SH), stadium (STA), storage tank (STO), tennis court (TC), train station (TS), vehicle (VE), and windmill (WM). It contains 23 463 images with 192 518 instances, including 11 725 images for training and 11 738 images for testing.

To evaluate the effectiveness of our method, two standard evaluation metrics, namely, mean Average Precision (mAP) and Correct Localization (CorLoc), are employed to evaluate the performance of WSOD in RSIs. Here, mAP is the evaluation metric to test the detection accuracy of object detectors on the testing set, and CorLoc is evaluated by measuring the localization accuracy on the training set [51]. Both two metrics employ the same IoU threshold of 0.5.

TABLE I
Ablation Studies of Proposal Generation on the NWPU VHR-10.v2 Test Set With the WSOD Frameworks of OICR and MELM

Proposal generation	WSOD frameworks			
	OICR		MELM	
Selective search	✓		✓	
SPG		✓		✓
mAP(%)	34.52	50.38	42.29	62.80

B. Implementation Details

In our experiments, we use a single RTX 2080Ti GPU with a batch size of 2 for the stochastic gradient descent (SGD) optimizer. The model iterates 20 epochs where the learning rate is 0.001. The momentum and the weight decay are set to 0.9 and 0.0005, respectively. We select VGG16 [52] pretrained on ImageNet [53] as the backbone network. We replace the spatial pooling layer after the last convolution layer with the RoI-pooling layer as [45]. For data augmentation, we resize images into five scales $\{480, 576, 688, 864, 1200\}$ (resize the shortest side to one of these scales), and each image is also augmented horizontally mirroring. The parameter setting of OICR and MELM follows the works of [26] and [27].

C. Ablation Experiments

The core contribution of this article is the proposed SPG method. This section conducts comprehensive ablation studies to evaluate the contributions of different components of our SPG method on the NWPU VHR-10.v2 dataset. For all the experiments, we use the same parameter settings.

First, to evaluate the contribution of our proposal generation approach, we compare the detection results of two proposal generation methods (our SPG and the popular SS method) with two basis WSOD frameworks (OICR [26] and MELM [27]) on the NWPU VHR-10.v2 test set. We first count the average numbers of the proposals obtained by SS and our SPG method on the NWPU VHR-10.v2 dataset. Specifically, SS [43] generates about 2000 proposals per image, and our SPG method introduces about 1100 proposals per image. Table I reports the comparison results. It can be seen that our proposed SPG method can effectively drive both the two WSOD frameworks to obtain better detection accuracy measured in terms of mAP. To be specific, our approach significantly outperforms the SS method by 15.86% mAP and 20.51% mAP, respectively, with the WSOD frameworks of OICR [26] and MELM [27]. We attribute the reasons for the significant improvement to that the proposals obtained by SS usually bring about redundant patterns, e.g., object parts and backgrounds, which causes localization randomness and model ambiguity.

In addition, in order to investigate the efficiency of our SPG method for generating object proposals, we conduct ablation studies by reporting the running times of different methods, respectively, on the NWPU VHR-10.v2 test set. Table II presents the comparisons of our method with two baseline methods in terms of the average time cost per image, which are tested with an NVIDIA 2080Ti GPU. It can be seen that the proposed SPG module can generate object proposals in

TABLE II

RUNNING TIME AMONG DIFFERENT METHODS ON THE NWPU VHR-10.v2 TEST SET

Methods	Proposal generation	Time(ms)
OICR [26]	Selective search	4200
MELM [27]	Selective search	3000
SPG+ OICR	SPG	600
SPG+ MELM	SPG	400

TABLE III

ABLATION STUDIES OF FEATURE ENHANCEMENT ON THE NWPU VHR-10.v2 TEST SET WITH THE WSOD FRAMEWORKS OF OICR AND MELM

Feature enhancement	WSOD frameworks			
	OICR		MELM	
Original feature maps	✓		✓	
Enhanced feature maps		✓		✓
mAP(%)	48.60	50.38	56.42	62.80

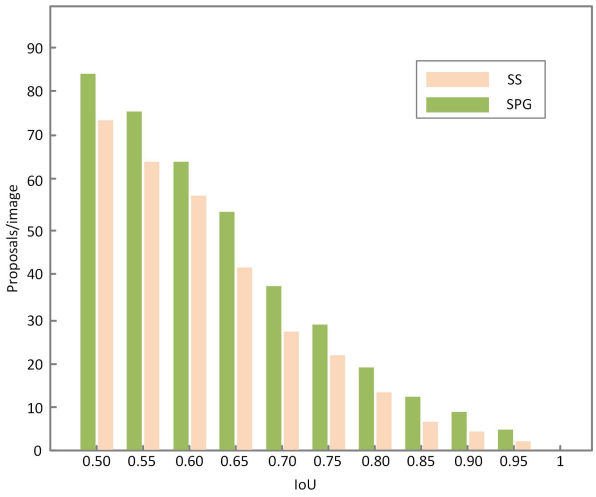


Fig. 3. IoU distribution of the proposals obtained by SS [43] and our SPG method on the NWPU VHR-10.v2 dataset.

less running time but with much higher detection accuracy (see Tables II and IV).

Meanwhile, in order to further verify the quality of the proposals obtained by our SPG method, we conduct the ablation study about the IoU distribution of the proposals obtained by two methods on the NWPU VHR-10.v2 dataset. Fig. 3 gives the results. As we can see, our SPG method obtains more proposals with high IoU than the SS method [43]. These promising results suggest that the quality of the proposals obtained by our SPG method is better than that of SS [43]. This is because our SPG method tends to precisely localize object regions after the confidence propagation procedure.

Besides, to demonstrate the validity of aggregating discriminative feature representation, we conduct another ablation study by using the original feature maps (without objectness confidence map) and the enhanced feature maps (with objectness confidence map for feature enhancement), respectively. Table III shows the experimental results. It is obvious that the enhanced feature maps can improve the detection accuracy from 48.60% mAP to 50.38% mAP with OICR [26] and from 56.42% mAP to 62.80% mAP with MELM [27], respectively. This demonstrates that enhancing the spatial information of convolutional features with the objectness confidence map can also remarkably boost detection accuracy.

D. Comparisons With State of the Arts

Here, we give some comparative experiments between our SPG method on the WSOD frameworks of OICR [26] and MELM [27] frameworks with other object detection methods on two challenging datasets.

1) Experimental Results on the NWPU VHR-10.v2 Dataset: Table IV shows the detection results for each class of 15 different methods, measured in terms of AP, on the NWPU-VHR-10.v2 test set. The upper part of the table lists the results of six fully supervised object detection methods, and then, nine weakly supervised methods are followed in sharp contrast to fully supervised object detection. 1) Our approach achieves the highest accuracy among all WSOD methods in terms of mAP. We have the following observations. In particular, our SPG method (with MELM) brings about the mAP improvements of 27.68%, 28.28%, 20.51%, and 23.39% compared with WSDDN [25], OICR [26], MELM [27], and PCL [28], respectively. Also, our SPG method (with MELM) is higher than the strongest competitor (PCIR [18]) by 7.83% mAP. 2) Build upon the OICR and MELM baselines, we get 15.86% and 20.51% improvements in terms of mAP, respectively. 3) Our method achieves comparable accuracy to some fully supervised object detection methods. Especially, our SPG method (with MELM) even surpasses the transferred CNN method. The benefits are mainly from our proposed SPG method. It could generate higher quality proposals than other comparison methods, all of which use SS to obtain object proposals. 4) Although our proposed method has boosted the accuracy with big margins, it is still difficult for the detection of the object classes of bridge, tennis court, and vehicle.

Table V indicates the localization precision comparisons among different methods on the NWPU VHR-10.v2 trainval set. It can be seen that our approach brings about the improvements of 38.17%, 33.40%, 23.54%, and 28.35%, measured in terms of CorLoc, compared with WSDDN [25], OICR [26], MELM [27], and PCL [28], respectively. The great improvement in localization precision proves the superiority of our SPG method once again.

Qualitative visualizations for both successful and failure examples on the NWPU VHR-10.v2 dataset are shown in Fig. 4. The results are demonstrated by rectangles together with the category labels in different colors. Here, green rectangles indicate success cases, while missed cases and false positives are highlighted in red and blue, respectively. Encouragingly, from the examples in Fig. 4, we can observe that most objects can be accurately and tightly covered by the predicted bounding boxes. Meanwhile, the instances that appear in adjacent locations have been accurately distinguished, which further verifies the effectiveness of our approach. Nevertheless,

TABLE IV

AVERAGE PRECISION COMPARISONS AMONG DIFFERENT 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 Object Detection											
Transferred CNN [1]	66.03	57.13	85.01	80.93	35.11	45.52	79.37	62.57	43.17	41.27	59.61
RICNN [3]	88.71	78.34	86.33	89.09	42.33	56.85	87.72	67.47	62.31	72.01	73.11
RCNN [54]	85.37	88.88	62.78	19.73	90.66	58.23	67.95	79.87	54.22	49.92	65.76
Fast RCNN [45]	90.91	90.60	89.29	47.32	100.00	85.85	84.86	88.22	80.29	69.84	82.71
Faster RCNN [55]	90.90	86.30	90.53	98.24	89.72	69.64	100.00	80.11	61.49	78.14	84.51
RICO [6]	99.70	90.80	90.61	92.91	90.29	80.13	90.81	80.29	68.53	87.14	87.12
Weakly Supervised Object Detection											
WSDDN [25]	30.08	41.72	34.98	88.90	12.86	23.85	99.43	13.94	1.92	3.60	35.12
OICR [26]	13.66	67.35	57.16	55.16	13.64	39.66	92.80	0.23	1.84	3.73	34.52
PCL [28]	26.00	63.76	2.50	89.80	**64.45**	76.07	77.94	0.00	1.30	15.67	39.41
MELM [27]	80.86	69.30	10.48	90.17	12.84	20.14	99.17	17.10	14.17	8.68	42.29
DCL [17]	72.70	74.25	37.05	82.64	36.88	42.27	83.95	39.57	16.82	35.00	52.11
PCIR [18]	**90.78**	78.81	36.40	90.80	22.64	52.16	88.51	42.36	11.74	35.49	54.97
DPLG [20]	80.90	78.30	10.50	90.10	64.40	**69.10**	80.20	39.60	14.00	8.70	53.60
SPG+OICR	42.80	72.56	58.82	81.24	28.78	50.25	96.78	35.60	12.24	24.75	50.38
SPG+MELM	90.42	**81.00**	**59.53**	**92.31**	35.64	51.44	**99.92**	**58.71**	**16.99**	**42.99**	**62.80**

TABLE V

LOCALIZATION PRECISION COMPARISONS AMONG DIFFERENT METHODS ON THE NWPU VHR-10.v2 TRAINVAL SET

Methods	Airplane	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Field	Harbor	Bridge	Vehicle	CorLoc(%)
WSDDN [25]	22.32	36.81	39.95	92.48	17.96	24.24	99.26	14.83	1.69	2.89	35.24
OICR [26]	29.41	83.33	20.51	81.76	40.85	32.08	86.60	7.41	3.70	14.44	40.01
PCL [28]	11.76	50.00	12.82	98.65	**84.51**	77.36	90.72	0.00	9.26	15.56	45.06
MELM [27]	85.96	77.42	21.43	98.33	10.71	43.48	95.00	40.00	11.76	14.63	49.87
PCIR [18]	**100.00**	**93.06**	64.10	99.32	64.79	79.25	89.69	62.96	13.26	52.22	71.87
DPLG [20]	87.20	85.10	16.80	96.10	75.10	73.20	86.30	46.70	**18.70**	16.30	61.50
SPG+OICR	44.73	88.29	45.02	92.80	45.05	58.24	93.70	25.60	8.56	27.00	52.90
SPG+MELM	98.06	92.67	**70.08**	**99.65**	51.86	**80.12**	**96.20**	**72.44**	12.99	**60.02**	**73.41**

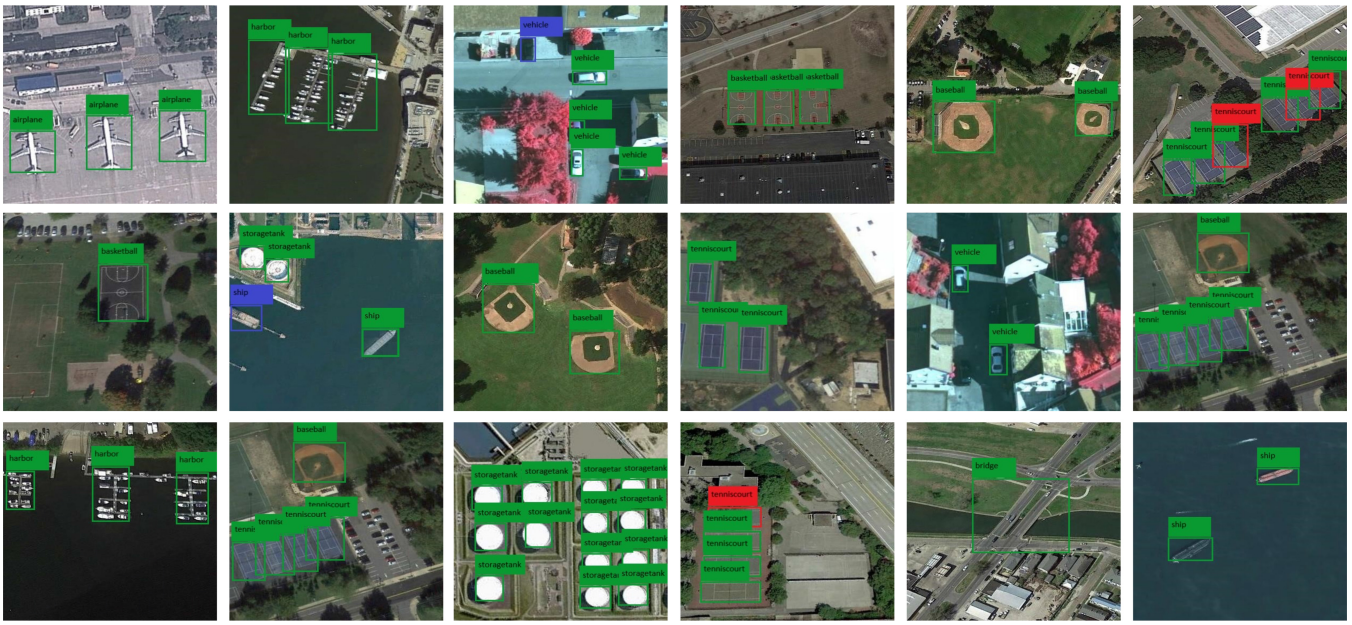


Fig. 4. Visualization of detection results on the NWPU VHR-10.v2 test split. Green bounding boxes indicate corrected cases, while missed cases and false positives are highlighted in red and blue, respectively.

TABLE VI

AVERAGE PRECISION COMPARISONS AMONG DIFFERENT METHODS ON THE DIOR TEST SET

Methods	APL	APO	BF	BC	BR	CH	DAM	ETS	ESA	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP(%)
WSDDN [25]	9.06	39.68	37.81	20.16	0.25	12.18	0.57	0.65	11.88	4.90	42.35	4.66	1.06	0.70	63.03	3.95	6.06	0.51	4.55	1.14	13.26
OICR [26]	8.70	28.26	44.05	18.22	1.30	20.15	0.09	0.65	29.89	13.80	57.39	10.66	11.06	9.09	59.29	7.10	0.68	0.14	9.09	0.41	16.50
PCL [28]	21.52	35.19	59.80	23.49	2.95	43.71	0.12	0.90	1.49	2.88	56.36	16.76	11.05	9.09	57.62	9.09	2.47	0.12	4.55	4.55	18.19
MELM [27]	28.14	3.23	62.51	28.72	0.06	62.51	0.21	13.09	28.39	15.15	41.05	26.12	0.43	9.09	8.58	15.02	20.57	9.81	0.04	0.53	18.66
DCL [17]	20.89	22.70	54.21	11.50	6.03	61.01	0.09	1.07	31.01	30.87	56.45	5.05	2.65	9.09	63.65	9.09	10.36	0.02	7.27	0.79	20.19
PCIR [18]	30.37	36.06	54.22	26.60	**9.09**	58.59	0.22	9.65	**36.18**	32.59	58.51	8.60	**21.63**	**12.09**	**64.28**	9.09	13.62	0.30	9.09	**7.52**	24.92
SPG+OICR	12.80	**38.27**	45.69	20.03	4.05	26.70	0.20	9.60	31.25	26.83	**62.73**	14.76	19.29	9.23	61.34	1.19	9.20	0.17	**9.12**	0.56	20.15
SPG+MELM	**31.32**	36.66	**62.79**	**29.10**	6.08	**62.66**	**0.31**	**15.00**	30.10	**35.00**	48.02	**27.11**	12.00	10.02	60.04	**15.10**	**21.00**	**9.92**	3.15	0.06	**25.77**

we can also observe that our proposed method gets some unsatisfactory detection results on the object categories of the tennis court with some missed instances on the NWPU VHR-10.v2 test set.

2) Experimental Results on the DIOR Dataset: Tables VI and VII indicate the detailed results for each class of several different WSOD methods, measured in terms of mAP and CorLoc, respectively, on the DIOR dataset. As seen

TABLE VII

LOCALIZATION PRECISION COMPARISONS AMONG DIFFERENT METHODS ON THE DIOR TRAINVAL SET

Methods	APL	APO	BF	BC	BR	CH	DAM	ETS	ESA	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	CorLoc(%)
WSDDN [25]	5.72	**59.88**	94.24	55.94	4.92	23.40	1.03	6.79	44.52	12.75	89.90	5.45	10.00	22.96	98.54	79.61	15.06	3.45	11.56	3.22	32.44
OICR [26]	15.98	51.45	94.77	55.79	3.55	23.89	0.00	4.82	56.68	22.42	91.41	18.18	18.70	31.80	98.28	81.29	7.45	1.22	15.83	1.98	34.77
PCL [28]	61.14	46.86	95.39	63.61	7.32	95.07	0.21	5.71	5.14	50.77	89.39	42.12	19.78	**37.94**	97.93	80.65	13.77	0.20	10.50	**6.94**	41.52
MELM [27]	76.98	28.94	92.66	63.01	13.00	90.09	0.21	16.96	37.88	44.62	88.08	49.39	15.65	28.19	98.28	82.97	22.75	10.34	4.62	2.23	43.34
PCIR [18]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.12
SPG+OICR	35.06	52.08	93.46	58.12	5.29	47.20	15.00	8.92	**58.43**	33.67	**92.99**	25.28	20.01	33.65	98.32	82.45	27.39	1.98	**16.22**	2.19	39.64
SPG+MELM	**80.48**	32.04	**98.68**	**65.00**	**15.20**	**96.08**	**22.52**	**16.99**	46.08	**50.96**	89.18	**49.45**	**22.00**	35.16	**98.61**	**90.04**	**32.56**	**12.73**	9.98	2.34	**48.30**

Fig. 5. Visualization of detection results on the DIOR test split. Green bounding boxes indicate corrected cases, while missed cases and false positives are highlighted in red and blue, respectively.

from Tables VI and VII, compared with the baseline method OICR [26], our SPG method (with OICR) obtains 3.65% mAP gains. We can also observe that our SPG method (with MELM) outperforms the baseline method MELM [27] by 7.11% mAP. Compared with the previous best-performing method PCIR [18] (24.92% mAP and 46.72% CorLoc) on the DIOR dataset, our proposed approach (25.77% mAP and 48.30% CorLoc) still acquires 0.85% mAP and 1.58% CorLoc gains. Thus, the detection performance on the DIOR dataset is also significantly boosted, which further demonstrates the validity of our proposed method. The relatively low detection accuracy and localization precision on the DIOR datasets are mainly because this dataset contains more complex scenarios and object categories.

The visualization of some detection results on the DIOR dataset is shown in Fig. 5. As seen in Fig. 5, most of the bounding boxes can perfectly enclose the objects, as shown with the green bounding boxes. However, our proposed approach also has trouble in addressing densely distributed objects and scene-ambiguous objects. This is mostly because of the following reasons: 1) there exist many densely distributed instances with the same category in RSIs, which misleads the object detectors to localize only one large bounding box and 2) due to the complexity of the background of RSIs and the lack of bounding box-level annotations, the background information often confuses object detectors to recognize the background as objects. These remain challenging issues, and we will consider introducing effective strategies (i.e., complementary attention learning) in the future.

V. CONCLUSION

In this article, we presented a simple yet surprisingly effective SPG method to boost the performance of WSOD in RSIs. To this end, the confidence propagation module is introduced to discover reliable locations of targets and generate the objectness confidence map. The high confidence regions within the objectness confidence map are utilized to progressively capture whole object locations. In addition, the objectness confidence map is further used to aggregate more powerful feature representatives. Extensive experimental results on the NWPU VHR-10.v2 and DIOR datasets verify that our proposed method significantly outperforms state-of-the-art methods on the WSOD task. Compared with the two baseline methods, our SPG achieves consistent improvement for each class by a large margin on the testing set. Generally, our proposed method outperforms the baselines OICR [26] and MELM [27] by 3.65% and 7.11% on the DIOR dataset, respectively, which are notable margins in terms of mAP. However, compared to fully supervised detectors, a large performance gap still exists for weakly supervised detectors. Our future work will focus on narrowing the performance gap.

REFERENCES

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

[2] G. Cheng, J. Han, L. Guo, and T. Liu, "Learning coarse-to-fine sparselets for efficient object detection and scene classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit*, Jun. 2015, pp. 1173–1181.

[3] 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.

[4] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 3–22, Nov. 2018.

[5] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[6] Z. Wu, W. Zhu, J. Chanussot, Y. Xu, and S. Osher, "Hyperspectral anomaly detection via global and local joint modeling of background," *IEEE Trans. Signal Process.*, vol. 67, no. 14, pp. 3858–3869, Jul. 2019.

[7] 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.

[8] 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.

[9] Z. Wu *et al.*, "Scheduling-guided automatic processing of massive hyperspectral image classification on cloud computing architectures," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3588–3601, Jul. 2020.

[10] 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.

[11] 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.

[12] W. Xie, J. Lei, S. Fang, Y. Li, X. Jia, and M. Li, "Dual feature extraction network for hyperspectral image analysis," *Pattern Recognit.*, vol. 118, Apr. 2021, Art. no. 107992.

[13] Z. Wu, J. Sun, Y. Zhang, Z. Wei, and J. Chanussot, "Recent developments in parallel and distributed computing for remotely sensed big data processing," *Proc. IEEE*, vol. 109, no. 8, pp. 1282–1305, Aug. 2021.

[14] 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.

[15] 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, 2016.

[16] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, "Weakly supervised learning for target detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 701–705, Apr. 2015.

[17] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 675–685, Jan. 2021.

[18] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8002–8012, Nov. 2020.

[19] X. Feng, J. Han, X. Yao, and G. Cheng, "TCANet: Triple context-aware network for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6946–6955, Aug. 2021.

[20] H. Wang *et al.*, "Dynamic pseudo-label generation for weakly supervised object detection in remote sensing images," *Remote Sens.*, vol. 13, no. 8, p. 1461, Apr. 2021.

[21] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 182–196, Dec. 2018.

[22] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112045.

[23] W. Xie, X. Zhang, Y. Li, J. Lei, J. Li, and Q. Du, "Weakly supervised low-rank representation for hyperspectral anomaly detection," *IEEE Trans. Cybern.*, vol. 51, no. 8, pp. 3889–3900, Aug. 2021.

[24] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, Jan. 1997.

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

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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. 8371–8380.

[32] G. Cheng *et al.*, "High-quality proposals for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 5794–5804, 2020.

[33] C. Lin, S. Wang, D. Xu, Y. Lu, and W. Zhang, "Object instance mining for weakly supervised object detection," in *Proc. Conf. Assoc. Adv. Artif. Intell*, vol. 34, no. 7, 2020, pp. 11482–11489.

[34] 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 Int. Conf. Comput. Vis. Pattern Recognit*, Jun. 2019, pp. 2194–2203.

[35] 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.

[36] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep self-taught learning for weakly supervised object localization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit*, Jul. 2017, pp. 4294–4302.

[37] C. Li, K. Yao, J. Wang, B. Diao, Y. Xu, and Q. Zhang, "Interpretable generative adversarial networks," in *Proc. IEEE Int. Conf. AAAI Conf. Artif. Int.*, 2022, pp. 1–9.

[38] G. Yan *et al.*, "C-MIDN: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9833–9842.

[39] 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.

[40] 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.

[41] K. K. Singh and Y. J. Lee, "You reap what you sow: Using videos to generate high precision object proposals for weakly-supervised object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit*, Jun. 2019, pp. 9406–9414.

[42] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 20, 2021, doi: 10.1109/TPAMI.2021.3074313.

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

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

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

[46] F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan, and X. Kong, "Random walks: A review of algorithms and applications," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 2, pp. 95–107, Apr. 2020.

[47] P. Vernaza and M. Chandraker, "Learning random-walk label propagation for weakly-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2953–2961.
[48] G. Bertasius, L. Torresani, S. X. Yu, and J. Shi, "Convolutional random walk networks for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2017, pp. 858–866.
[49] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit*, Jun. 2018, pp. 4981–4990.
[50] M. E. Newman, "The mathematics of networks," *New Palgrave Encyclopedia Econ.*, vol. 2, pp. 1–12, Sep. 2008.
[51] 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.
[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[53] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[54] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit*, Jun. 2014, pp. 580–587.
[55] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

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 currently a Professor with Northwestern Polytechnical University. His main research interests are computer vision, pattern recognition, and remote sensing image understanding.

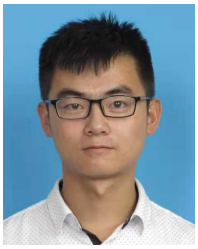
Dr. Cheng is also an Associate Editor of *IEEE Geoscience and Remote Sensing Magazine* and a Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Xuan Xie received the master's degree from Shenyang Aerospace University, Shenyang, China, in 2019. She is currently pursuing the Ph.D. degree with Northwestern Polytechnical University, Xi'an, China.

Her research interests include computer vision and image processing, especially remote sensing image analysis.

Weining Chen received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree with Northwestern Polytechnical University, Xi'an.

His research interests include aerial remote sensing image acquisition and processing.



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

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

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

He is currently an Associate Professor with Northwestern Polytechnical University. His research interests include computer vision and remote sensing image processing, especially in fine-grained image classification and object detection.

Junwei Han (Fellow, 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 currently a Professor with Northwestern Polytechnical University. His research interests include computer vision and brain imaging analysis.