Multiple Instance Graph Learning for Weakly Supervised Remote Sensing Object Detection

Binglu Wang, *Member, IEEE*, Yongqiang Zhao, *Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*— Weakly supervised object detection (WSOD) has recently attracted much attention in the field of remote sensing, where only image-level labels that distinguish the existence of an object in images are required. However, existing methods frequently treat the most discriminative area of an object as the optimal solution and, meanwhile, ignore the fact that more than one instance may exist in a certain class in remote sensing images (RSIs). To address the issue, we propose a unique multiple instance graph (MIG) learning framework for WSOD in RSIs. The motivation of this work is twofold: 1) a spatial graph-based vote (SGV) mechanism is proposed to find high-quality objects by collecting the top-ranking votes with highly spatial overlap and 2) an appearance graph-based instance mining (AGIM) model is further constructed to exploit all possible instances with the same class by propagating the label information according to the apparent similarity. It is noted that the formulated MIG framework that collaborates SGV and AGIM is independent of extra hyperparameters or annotations. Experimental results reported for two well-known benchmarks, i.e., NWPU VHR-10.v2 and DIOR, testify to the superiority of the proposed framework by 55.9% and 25.11% mAPs.**

***Index Terms*— Multiple instance graph (MIG) learning, object detection, remote sensing images (RSIs), weakly supervised learning.**

I. Introduction

OBJECT detection focuses on simultaneously localizing and recognizing object instances in given images. It is a fundamental technique in analyzing remote sensing images (RSIs) [1]–[6]. Currently, breakthrough progress on object detection has been boosted by the development of powerful convolutional neural networks (CNNs) [7], [8] and the techniques of spatial resolution enhancements [9]. However, these predominated works rely on large-scale datasets with subtle manual annotations [10], [11]. It is time-consuming and labor-intensive to collect fine-grained annotations large-scale RSIs, thus significantly restricting the wide execution of object detection technologies in real applications. To alleviate the heavy labeling cost, weakly supervised learning has been extensively explored. It utilizes only uncompleted image-level labels that declare whether an object category appears in images. As a result, this article aims at learning precise object detectors by leveraging weakly supervised learning techniques. Most previous methods [12]–[16] resort to multiple instance learning (MIL) for weakly supervised object detection (WSOD) problems via dividing region proposals into positive and negative bags. A positive bag contains at least one object instance, and none of the instances could appear in a negative bag. Among them, a weakly supervised deep detection network (WSDDN) [12] first integrated MIL into the WSOD model and formulated the end-to-end object detection model under weakly supervised settings. After that, Tang *et al.* [17] introduced a novel online instance classifier refinement (OICR) strategy implemented by multiple stages in a single deep network. It is an impressive work and facilitates a series of advanced WSOD works. Based on it, recent approaches [18]–[24] use context information [18], [22], [24], segmentation [19], or learning strategies [20], [21], [23] to empirically regularize the learning procedure and achieved remarkable performance [25].

Nevertheless, WSOD in RSIs still encounters two challenges. First, existing weakly supervised methods usually tend to converge the most contributing areas and the related background that results from the regions usually contribute more to classification results. Even though the problem exists in both nature images and RSIs, it leads to worse detection performance in RSIs as they usually contain more foreground-related context regions. For example, ships are usually close to water, and bridges are usually above rivers. It is the main reason why the performance of WSOD is lower than methods with full supervision in RSIs. Second, previous WSOD methods usually ignore the diversity within a class in RSIs: there is usually more than one instance that appears for the same class in an RSI (e.g., basketball court, airplane, and tennis court), while previous methods with weak supervision intend to simply select only one discriminative region to train the detector, which results in suboptimal detectors. It is another important issue that significantly limits the development of WSOD in RSIs.

To address the first issue, we draw on the training strategy of fully supervised object detection [26]–[28]. The ground-truth

Manuscript received June 10, 2021; revised August 27, 2021 and October 13, 2021; accepted October 20, 2021. Date of publication October 26, 2021; date of current version February 15, 2022. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61771391, in part by the Science, Technology, and Innovation Commission of Shenzhen Municipality under Grant JCYJ20170815162956949 and Grant JCYJ20180306171146740, in part by the Key Research and Development Plan of Shaanxi Province under Grant 2020ZDLGY07-11, and in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2018JM6056. *(Corresponding author: Yongqiang Zhao.)*

Binglu Wang is with the School of Automation Engineering and the School of Artificial Intelligence, Optics, and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: wbl921129@gmail.com).

Yongqiang Zhao is with the School of Automation Engineering, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Research & Development Institute, Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China (e-mail: zhaoyq@nwpu.edu.cn).

Xuelong Li is with the School of Artificial Intelligence, Optics, and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: li@nwpu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3123231

bounding boxes are treated as cluster centers, and then, the same label is assigned to the proposals closely surrounding it. However, the instance-level bounding boxes are unavailable in weakly supervised learning. We observe that the top-ranking proposals can roughly diagnose the localization of objects as they usually cover the discriminative parts of objects. Meanwhile, highly overlapped proposals naturally belong to the same label. Thus, propagating label information within an undistinguished image can help to discover more object extent. To this end, we propose a spatial graph-based vote (SGV) strategy to highlight the high-quality object. Specifically, we first collect top-ranked proposals by virtue of the K-means algorithm. The proposal scores are split into different clusters, and the cluster with the highest score is selected. Then, we construct an undirected spatial graph according to the spatial similarity of the obtained proposals, where two proposals with large spatial overlaps are connected. In terms of discovering high-quality objects, we further find that high-quality objects should have the highest overlaps with others in the graph. Therefore, selecting the instance that possesses the most spatial votes than others can mine full object extent. Each box with a high overlap proposal will get a spatial vote.

To address the second issue, we further propose an appearance graph-based instance mining (AGIM) model based on a fundamental assumption: the instances from the same class should have a corresponding apparent similarity. Formally, we greedily collect same-class instances by capturing all possible instances that obtain similar appearance similarities with the most confident proposal. Meanwhile, we treat implicit appearance similarities existing in the built spatial graph as a metric to mine inconspicuous instances of the same class. None of the extra hyperparameters are introduced in the AGIM module. By collaborating SGV with AGIM, a novel and flexible multiple instance graph (MIG) learning framework is formulated. SGV is used to pursue high-quality instances and supplies implicit bonds. AGIM captures more abundant intraclass information, which facilitates a more robust object detector. We clarify the main contributions of this work as follows.

1) We propose an MIG learning framework for WSOD in RSIs, which makes the object detector covers a more complete object area and is capable of detecting multiple objects in the same class.
2) We design a spatial graph-based vote mechanism to pursue high-quality objects by seeking objects who have that have high spatial overlap with top-ranked proposals.
3) We introduce a parameter-free appearance graph instance mine strategy to flexibly mine all possible instances from the same class under weakly supervised settings.
4) Experimental results on two public datasets, NWPU VHR-10.v2 and DIOR, demonstrate the effectiveness of the proposed MIG learning framework.

II. Related Work

A. Fully Supervised Object Detection

In the past decade, object detection problems have been extensively studied for both nature images and RSIs [29]–[33]. With the boom of CNNs [6]–[8], [34], [35] and the availability of large-scale datasets equipped with manual subtle annotations, lots of representative researches [36]–[43] have emerged in the field of object detection and achieved impressive progress. For instance, Girshick *et al.* [27] proposed a breakthrough Fast R-CNN and inspired a lot of works. Cheng *et al.* [44] introduced a rotation-invariant layer and effectively alleviated the problem of versatile angles in geographic images. Tang *et al.* [45] proposed a hyper region proposal network (HRPN) combined with a cascade of boosted classifiers to detect vehicles in RSIs. Yang *et al.* [46] constructed a Markov random field (MRF) fully convolutional network to detect airplanes. Huang *et al.* [29] propose a nonlocal-aware pyramid attention module to make the detector suppress background noise and introduce a multiscale refinement feature pyramid module to choose the optimal receptive field. Tang *et al.* [32] introduce a graph attention network into the remote sensing object detection. Oriented object detection has also been explored in RSIs [31], [33]. These methods are data-driven and rely on precise instance-level labels.

B. Weakly Supervised Object Detection

Weakly supervised learning has attracted widespread research attention since it can greatly alleviate the heavy labeling burden. The technique has been well studied in image [47], [48] and video recognition [49]. WSOD [50]–[53] is one of the most popular tasks in weakly supervised learning due to its potential value in many applications, such as lesion detection and remote sensing retrieval. Current works [17], [20], [50], [54]–[58] solve WSOD problems with a two-stage approach, i.e., the object proposal methods are first leveraged to decomposed images into a series of region proposals. Then, the WSOD task is simplified as a multilabel problem by combining MIL learning, which divides region proposals into positive and negative bags. Following MIL constraints, the main task of WSOD is to find representative object instances from positive bags to train detectors. Unfortunately, in the process, the most representative part is usually selected as the positive example rather than the whole object, which is one of the main reasons triggering the inferior results of WSOD. Meanwhile, MIL strategy leads to a nonconvex optimization problem. On this occasion, the model is sensitive to positive instance initialization and easily tends to get stuck in local extrema.

In nature image-based WSOD, many efforts attempt to overcome these issues via creating better initializations or improving the optimization strategy. For example, Tang *et al.* [17] proposed an OICR algorithm to mitigate the local extrema problem. Bilen and Vedaldi [12] presented an end-to-end WSDDN based on the framework of Fast R-CNN, which uses the product of the proposal objectness score and spatial recognition score as a standard to select the positive example. Wang *et al.* [59] combined the min-entropy latent model with the WSOD framework and successfully boosted the performance of WSOD. Although the aforementioned methods have achieved progressive performance in natural images, they tend to perform poorly in RSIs as its large-scale cluttered background. Meanwhile, these methods also ignore the fact

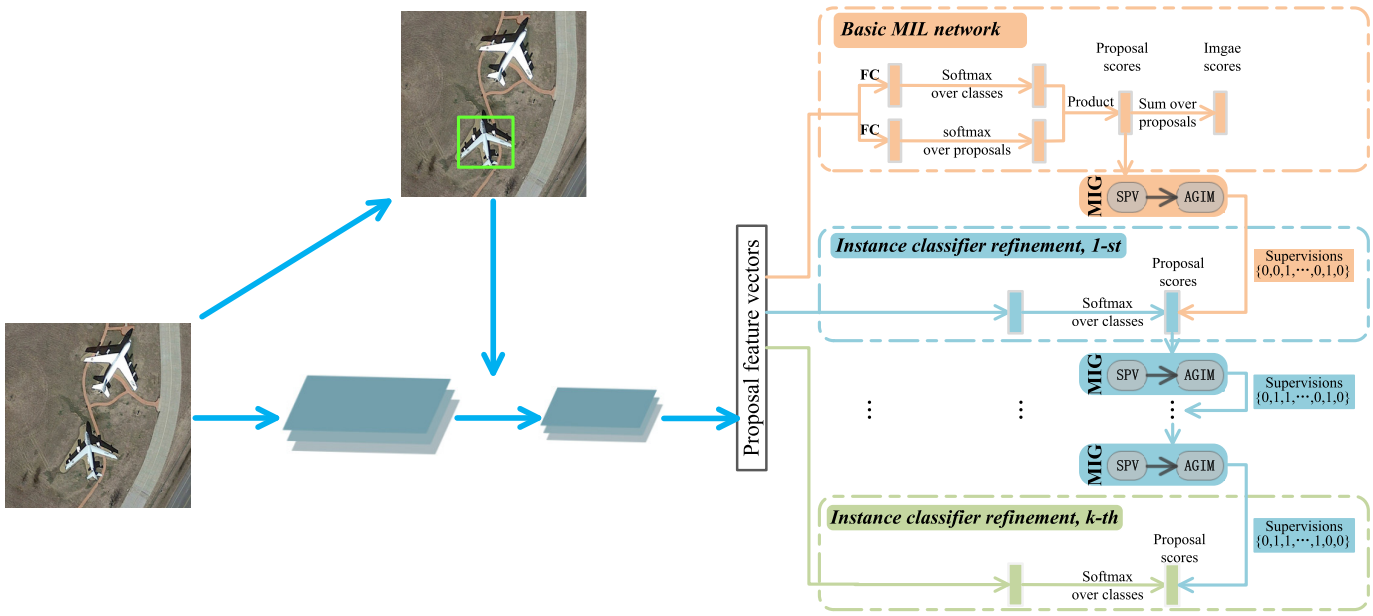

Fig. 1. Overview of the proposed MIG learning framework. The input image and its region proposals are fed into the backbone network to obtain corresponding proposal features by employing the RoI pooling. Then, the obtained proposal features are branched into many paralleled streams: the first steam is basic MIL network and the others are instance classifier refinement. In every stage, the proposed MIG is leveraged to mine all possible instances according to the scores of proposals from the former branch and provides supervision for their latter streams.

that more than one instance from the same class appears in an RSI.

In recent years, many works have been proposed to solve the problem of WSOD in RSIs. Han *et al.* [60] iteratively trained the detector with refined annotations to make the model converge. Yao *et al.* [23] introduced the curriculum learning strategy into the training of the object detectors to obtain a robust detection model. Zhou *et al.* [61] proposed a negative bootstrapping-based model and utilized transferred deep features. Feng *et al.* [22] introduced a progressive contextual instance refinement to pursue high-quality instances. The work [24] constructed a triple context-aware network to better detect multiple adjacent objects in RSIs. Furthermore, many methods devoted to resolving the rotation problem in remote sensing detection [11], [36], [44], [62]–[64]. Li *et al.* [11] introduced an additional region proposal network (RPN) to capture the multiangle and multiscale characteristics. Cheng *et al.* [44] proposed a rotation-invariant CNN (RICNN) layer to enforce the rotated object features to be mapped closely to unrotated ones. Duan *et al.* [62] constructed a rotation-invariant local binary descriptor where all the rotation variants of a patch are rotated into the same orientation and are projected into the same binary descriptor. In this article, we pay attention to mine full object extent and, meanwhile, discover all possible same-class instances in given images under weakly supervised settings.

III. Proposed Method

A. Preliminaries

Bilen and Vedaldi [12] are among the first to integrate MIL into WSOD task and formulate an end-to-end framework. It inspires a lot of WSOD works and is widely used to preliminarily localize the instances. Specifically, as shown in Fig. 1, given a training image I and its region proposals R, which are generated by [68] and [69], and image-level label $Y = [y_1, \ldots, y_c, \ldots, y_C] \in \{0, 1\}$, which indicates whether an object category exists in an image or not, a neural network pretrained on ImageNet with RoI pooling [27] is first employed to generate a set of fix-sized proposal features. Then, the obtained features are branched into two paralleled and fully connected layers to, respectively, generate classification logits $\Psi_{\text{cl}}(c, r) \in \mathbb{R}^{|R|\times C}$ and detection logits $\Psi_{\text{det}}(c, r) \in \mathbb{R}^{|R|\times C}$ for each object category and each region, where $|R|$ is the number of region proposals and C denotes the object category number. The classification score $s_{\text{cls}}(c, r)$ and the detection score $s_{\text{det}}(c, r)$ are generated by applying softmax operations along different directions

$$\begin{cases} s_{\text{cls}}(c, r) = \dfrac{\exp \Psi_{\text{cls}}(c, r)}{\sum_{c\in C} \Psi_{\text{cls}}(c, r)} \\ [3mm] s_{\text{det}}(c, r) = \dfrac{\exp \Psi_{\text{det}}(c, r)}{\sum_{r\in R} \Psi_{\text{det}}(c, r)} \end{cases} \tag{1}$$

where $s_{\text{cls}}(c, r)$ denotes the region r being classified as category c and $s_{\text{det}}(c, r)$ denotes the contribution of region r to image being classified to class c. The scores of proposal $s(c, r)$ are computed by an elementwise product: $s(c, r) = s_{\text{cls}}(c, r) \odot s_{\text{det}}(c, r)$. Finally, the image-level score is obtained via summing over all region scores: $s_c = \sum_{r\in R} s(c, r)$. During training, the WSDDN network is optimized by the following loss function:

$$\mathcal{L}_{\text{wsddn}} = -\sum_{c=1}^{C} \{y_c \log s_c + (1 - y_c) \log(1 - s_c)\}. \tag{2}$$

Meanwhile, multistage instance refinement branches with $C+1$ dimensions are integrated paralleled with aforementioned branches to alleviate the local-optimal problem, where $C + 1$

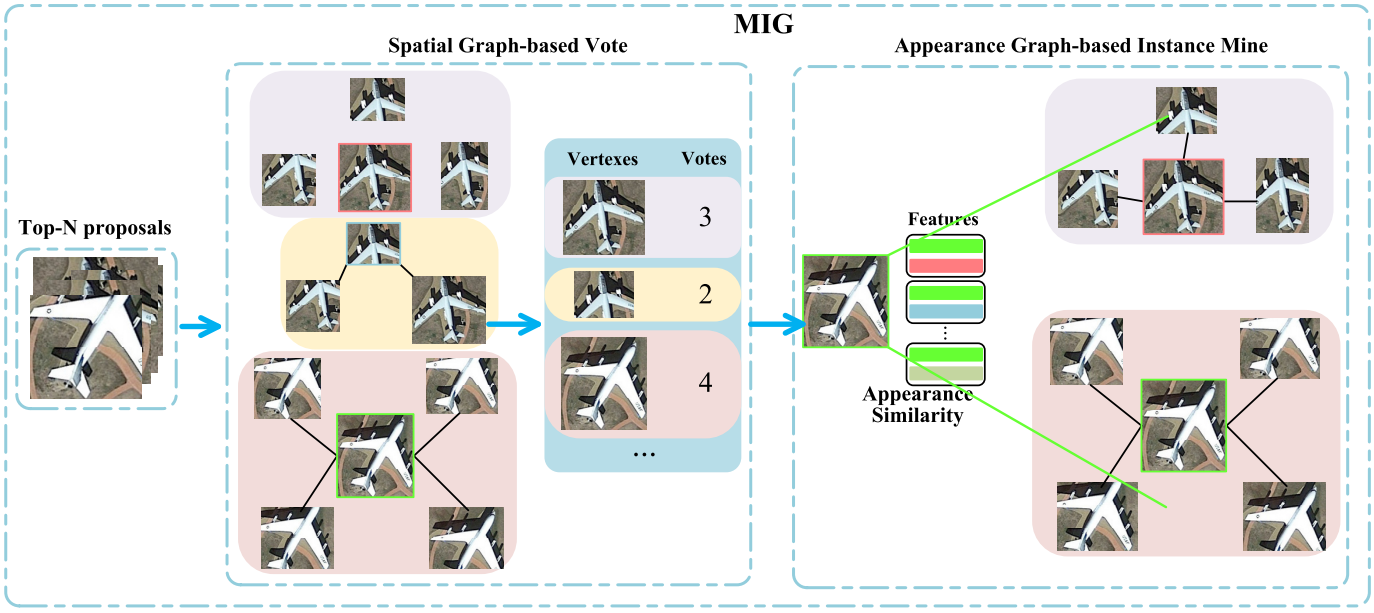

Fig. 2. Illustration of the proposed MIG method. To mine all possible high-quality instances in an image, MIG first collects top-n proposals via employing the K-means algorithm. Then, the spatial graph-based vote strategy is constructed according to the spatial similarity for the obtained top-ranking proposals, and the proposal that has the highest overlaps with others under the built graph is identified high-quality instance. Finally, the AGIM algorithm is further proposed according to their apparent similarity between the high-quality proposal and others.

denotes the number of object classes and one background. $s_k(c, r)$ is the corresponding detection scores for the kth instance refinement branch. During the training stage, the top-scoring proposal and its adjacent regions from the former branch are treated as pseudoinstance-level labels y_{cr}^k, which supervises the latter branch learning. The loss for the instance refinement stage is defined as

$$\mathcal{L}_r^k = -\frac{1}{R}\sum_{r=1}^{R}\sum_{c=1}^{C+1}\omega y_{cr}^k \log s_k(c, r) \tag{3}$$

where $s_k(c, r)$ denotes the score of proposal r in the kth refinement stage and ω is loss weights. In this article, we propose a novel MIG learning framework and embed it into the aforementioned WSOD framework to pursue all possible same-category instances without introducing extra hyperparameters and annotations.

B. Spatial Graph-Based Vote Strategy

Weakly supervised methods usually tend to converge the most contributing parts and their strongly related background as those regions usually contribute more to classification results. This problem becomes worse in RSIs as there are more foreground-related context regions in many remote sensing objects. Although leveraging multistage instance refinements can boost the detection performance to some extent, such progressive refinement operation is limited by the quality of the initial object proposal. In other words, if there are no reasonable proposals for model initialization, this refinement strategy cannot correctly discover the whole objects. It introduces a critical risk: the instance refinement process cannot discover the correct object when unreasonable object candidates are selected as pseudolabels. Thus, it is very important to mine the credible instance at the beginning of instance refinement.

In the training of fully supervised object detection, the ground-truth bounding boxes are always treated as cluster centers and then assign other proposals closely surrounding it to the same label. Inspired by that, we can draw the following conclusion that the ground-truth bounding boxes have the highest overlaps with others. Based on the above, although the top-scoring proposal may only cover the discriminative part, its surrounding proposals with suboptimal scores may capture the whole object. To address it, we introduce a novel spatial graph vote strategy to pursue high-quality objects by iteratively selecting the proposals that have enough scores and most spatial overlaps with others.

More specifically, as illustrated in Fig. 2, when an image has object class label c (i.e., $y_c = 1$), for the kth instance refinement, OICR [17] only selects the top-scoring proposal as positive instance to conduct instance refinement. However, the top-scoring proposals usually cover the low-quality parts of objects, which is the main reason for its inferior results. To address this problem, we first employ the K-means algorithm to divide the proposals into some clusters and then treat the highest score center as top-n proposals with indexes $D_c^k = \{r_{\mathrm{ck}}^1, \ldots, r_{\mathrm{ck}}^N\}$. Then, an undirected unweighted spatial graph $G_{\mathrm{ck}}^s = (V_{\mathrm{ck}}^s, E_{\mathrm{ck}}^s)$ is constructed for these proposals according to their spatial correlation. We define the top-n proposals as vertexes V_{cl}^s, and $E_{\mathrm{ck}}^s = \{\sigma_{\mathrm{ck}}^{rr'}\} = \{\sigma(v_{\mathrm{ck}}^r, v_{\mathrm{ck}}^{r'})\}$ as edges to denote the connections between the vertexes, which can be computed in

$$\sigma_{\mathrm{ck}}^{\mathrm{rr}'} = \begin{cases} 1, & I_{\mathrm{rr}'} > I_T \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where I_T is a threshold and $I_{\mathrm{rr}'}$ is the spatial correlation between the rth proposal R_r and the r'th proposal $R_{r'}$

$$I_{\mathrm{rr}'} = \frac{|R_r \cap R_{r'}|}{|R_r \cup R_{r'}|}. \tag{5}$$

Algorithm 1 Spatial Graph-Based Vote Strategy

Input: Refinement times K; Training images I_i; Region proposals R_i; Image-level label $Y = [y_1, \ldots, y_c, \ldots, y_C]$

```
Feed image and its proposals (I, R) into the network, obtain proposal score matrices s_k(c, r), k ∈ {0, ..., K − 1}
for k = 0 to K − 1 do
    for c = 1 to C do
        if y_c = 1 then
            Generate top-n proposals with indexes D_c^k.
            Construct an indirect-unweighted graph G_c^k.
            Compute spatial votes v_cr for each vertex.
            Obtain the credible positive instance P_c.
        end if
    end for
end for
```

Output: Positive instance P_c.

Accordingly, we greedily connect the vertexes if they have enough spatial overlaps for class c using this graph. According to the above analysis, the high-quality object should have the most connections. We identify an effective connection (i.e., $\sigma(v_{\mathrm{ck}}^{r}, v_{\mathrm{ck}}^{r'}) = 1$) as a positive spatial vote. The corresponding spatial votes for the vertex can be computed by

$$v_{\mathrm{cr}} = \sum_{r' \in R} \sigma_{\mathrm{ck}}^{\mathrm{rr}}. \tag{6}$$

Based on it, we can directly select the vertex that has the most spatial votes as a credible positive instance, such as the cluster center, is fully supervised settings, which is denoted as

$$P_c = \max v_{\mathrm{cr}}. \tag{7}$$

Then, all the nodes in the built spatial graph G_c^k will be labeled to the same class as P_c. Accordingly, we can get high-quality initialization proposals to conduct instance refinement, thereby alleviating the problem of part domination and reducing the critical risk. The process of SGV is shown in Algorithm 1. It is notable that the number of graphs depends on the number of categories in the image.

C. Appearance Graph-Based Instance Mining Algorithm

Due to the absence of instance-level annotations, the existing method only focuses on how to mine the most confident instance but ignores the fact that there are many same-class instances appearing in one RSI. Although the above strategy has successfully boosted the performance of WSOD, it not only seriously ignores the diversity of information within the class but also regards unexplored positive instances as negative samples during the training, thereby hurting the discrimination of object detectors. To address this issue, we further construct an AGIM strategy to mine all possible instances for the corresponding object detector learning.

Based on our observation, instances in RSIs belonging to the same category must appear in different spatial locations. We can draw a fundamental assumption: if different proposals have high apparent similarity and appear in different locations, they are different instances of the same category. Based on it, we can discover more possible instances of the same category by considering their apparent similarity and spatial correlation simultaneously.

Formally, as shown in Fig. 2, given a training image I_i and its region proposals R_i and their corresponding detection scores $s_k(c, r)$ in the kth instance refinement, let $\mathcal{F} = \{f_1, \ldots, f_R\}$ denote the feature vectors of proposals R_i, and we can obtain them on the fully connected layer. In Section III-B, we have obtained the credible instance by constructing an undirected unweighted spatial graph and labeling all nodes with the same class. Thus, we build an undirected unweighted spatial graph $G_{\mathrm{ck}}^{a} = (V_{\mathrm{ck}}^{a}, E_{\mathrm{ck}}^{a})$, where the proposals having apparent similarity with core instance $\mathcal{P}_c$ is denoted as vertexes V_{ck}^{a}, and E_{ck}^{a} is the edge to denote the corresponding apparent similarity. In this article, we employ the Euclidean distance to evaluate the apparent similarity between the core instance $\mathcal{P}_c$ and other proposals, which is given by

$$E_{\mathrm{ck}}^{a} = D_{i,\mathcal{P}_c} = \| f_i - f_{\mathcal{P}_c} \|_2. \tag{8}$$

To better evaluate the apparent similarity of the same class, we generate an interclass similarity threshold by computing the average distance of all the vertexes in G_{ck}^{a} and define it as

$$T = \frac{1}{N} \sum_{i} D_{i,\mathcal{P}_c}, \text{ s.t.} \frac{|\mathcal{P}_c \cap R_i|}{|\mathcal{P}_c \cup R_i|} > 0.5 \tag{9}$$

where N is the number of vertexes in G_{ck}^{s} and R_i denotes the vertexes that meet the constraints above. The proposals meet the condition that $D_{i,\mathcal{P}_c} < T$ is preliminarily identified as the same class instances. Next, the NMS with threshold 0.3 is applied to reduce the spatial duplicated proposals and obtain the foreground bounding boxes. Finally, we can get the supervision for training in the next stage. To be specific, for the rth proposal in the kth refinement stage, if it is regarded as the foreground bounding for class c, the supervision of this proposal in the $(k+1)$th refinement stage is set to 1 on class c, i.e., $y_{k+1}(c, r) = 1$; otherwise, it would be set as 0. Accordingly, we can discover all possible instances by constructing an apparent similarity but spatial-irrelevant graph. In such a way, more abundant intraclass information can be captured to facilitate the robust detector. The detailed process of AGIM is shown in Algorithm 2.

The proposed MIG module absorbs and learns the PCL [20] while initializing the graph, but there exist some essential differences. PCL aims to actually detect objects in nature images via proposal cluster. However, since PCL directly uses the top-ranking proposals to generate the supervision information, the number of objects that PCL can detect is highly close to the number of K-means clusters and makes PCL cannot deal with remote sensing object detection well. Different from PCL, our MIG module not only uses the top-ranking proposals as vertexes but also introduces proposals that have a similar appearance to the top-ranking proposals, and the edges are used to denote the apparent similarity between vertexes. This design can make the detector find more instances in one class, which is more in line with the characteristics of RSIs.

Algorithm 2 AGIM Algorithm

Input: Refinement times K; Training images I_i; Region proposals R_i; image-level label $Y = [y_1, \ldots, y_c, \ldots, y_C]$

Feed image I and its proposals into the network to produce proposal score matrices $\mathcal{F} = \{f_1, \ldots, f_R\}$
for $k = 0$ **to** $K - 1$ **do**
  for $c = 1$ **to** C **do**
    if $y_c = 1$ **then**
      $V^s_{\text{ck}} \leftarrow \varnothing, V^a_{\text{ck}} \leftarrow \varnothing, T \leftarrow 0$
      Obtain $\mathcal{P}_c$ by Algorithm 1.
      $V^s_{\text{ck}} \leftarrow \mathcal{P}_c, V^a_{\text{ck}} \leftarrow \mathcal{P}_c$
      Evaluate E^a_{ck} of core instance $\mathcal{P}_c$.
      Generate inter-class similarity threshold T.
      for N=1 **to** $|R|$ **do**
        if $D_{i,\mathcal{P}_c} < T$ **then**
          $V^a_{\text{ck}} \leftarrow R_i$
        end if
      end for
      Apply NMS to V^a_{ck}.
    end if
  end for
end for

Output: All vertexes V^a_{ck} in appearance graph G^a_{ck}.

IV. Experiment

In this section, we first provide detailed descriptions of our experiment settings (i.e., datasets, evaluation metrics, and implementation details). We conduct extensive ablation experiments to validate the contribution of each module. Comparisons with the state of the arts are further provided to demonstrate the efficacy of the proposed method.

A. Datasets and Evaluation Metrics

Following the current state-of-the-art WSOD works in RSIs, we first conduct experiments on the public commonly used NWPU VHR-10.v2 dataset [11], which consists of 1172 images with the size of 400 × 400 pixels and covers ten object categories. It is composed of three groups, including a training set with 679 images, a validation set with 200 images, and a testing set with 293 images. We also evaluate our approach on the larger and more challenging DIOR dataset [10], which consists of 23 463 images with the size of 600 × 600 pixels, includes a total of 192 472 instances, and covers 20 object categories, including airplane (C1), airport (C2), baseball field (C3), basketball court (C4), bridge (C5), chimney (C6), dam (C7), expressway service area (C8), expressway toll station (C9), golf field (C10), ground track field (C11), harbor (C12), overpass (C13), ship (C14), stadium (C15), storage tank (C16), tennis court (C17), train station (C18), vehicle (C19), and wind mill (C20).

As with all WSOD methods, we combine the training set with the validation set formulating a trainval set for WSOD training, and the testing set is employed for testing. On the one hand, the correct localization (CorLoc) is applied to evaluate the localization accuracy on the trainval set. On the other hand, the average precision (AP) is employed to evaluate the testing accuracy on the testing set. Both metrics employ the same IoU threshold with 0.5 to identify the result as a positive detection.

B. Implementation Details

For a fair comparison, we adopt VGG16 [68] pretrained on the ImageNet dataset [69] as the backbone network where the Pooling4 layer is removed and the Pooling5 layer is substituted with ROI pooling [27]. The settings for network training are also kept identical to previous WSOD methods [17], [22], [24], [59], including learning rate, minibatch, weight decay, and momentum. They are set to 0.001, 2, 0.005, and 0.9, respectively. During training, we perform 30k iterations with a 10k step size for the NWPU VHR-10.v2 dataset and 200k iterations with a 100k step size for the DIOR dataset. We also employ selective search to generate about 2000 proposals per image and take horizontal flipping along with random five image scales. During testing, NMS of 0.3 is used to reduce the duplicated bounding boxes. All the experiments are conducted on Ubuntu 16.04, NVIDIA GeForce RTX 2080, cuDNN v5, and CUDA 9.0.

C. Ablation Experiments

As shown in Table I, our ablation studies are performed on the DIOR dataset to ablate the contributions of each module, including the SGV and AGIM. The "+ SGV" denotes baseline with our SGV strategy, the "+ AGIM" denotes baseline with our AGIM algorithm, and "+ MIG" denotes baseline with our MIG method.

1) Baseline Setup: We select MELM [59] as our baseline for the ablation study. For a fair comparison, we keep all experimental settings strictly consistent.

2) Effect of Spatial Graph-Based Vote Strategy: In order to evaluate the contribution of SGV, we replace the OICR-based instance mining strategy, which only selects the top-scoring proposal as pseudoinstance-level label, with the proposed SGV strategy. As shown in Table I, leveraging SGV strategy brings a large improvement compared to baseline (i.e., mAP is boosted from 18.66% to 22.0% and CorLoc is improved from 43.34% to 45.19%). The results fully demonstrate the effectiveness of the proposed SGV. The reason for the improvement is given as follows: due to the nonconvex optimization process, the WSOD model inclines to discover discriminative parts so that top-scoring proposals usually cover the object part rather than the whole object, which leads to inferior results for baseline. Different from it, our SGV strategy aims to find a credible object cluster center, such as fully supervised ground-truth bounding boxes by collecting the top-ranking votes with highly spatial overlap. Accordingly, the more accurate and tight pseudoinstance-level label can be mined to train the corresponding detector, thereby alleviating the part domination problem in WSOD.

3) Effect of Appearance Graph-Based Instance Mining Algorithm: To testify the effectiveness of the proposed AGIM, we first integrate it into our baseline network. Note that we generate the interclass similarity threshold T in (9) by computing the average distance of the top-scoring proposal and its adjacent proposals. As clearly indicated in Table I, the detection performance can be significantly boosted where mAP is

TABLE I
RESULTS ON THE DIOR DATASET FOR EACH MODULE

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Results
	mAP										
Baseline	**0.2814**	0.0323	0.6251	0.2872	0.0006	0.6251	0.0021	0.1309	0.2839	0.1515	0.1866
+ SGV	0.1627	0.4566	0.5704	0.2631	0.0734	0.6353	0.0019	**0.3472**	0.2763	0.3764	0.2200
+ AGIM	0.2091	0.4265	0.6058	**0.3011**	0.0214	0.6390	**0.0080**	0.2202	0.2603	0.4624	0.2324
+ MIG	0.2220	**0.5257**	**0.6276**	0.2578	**0.0847**	**0.6742**	0.0066	0.0885	**0.2871**	**0.5728**	**0.2511**
Methods	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	Results
	mAP										
Baseline	0.4105	**0.2612**	0.0043	**0.0909**	0.0858	0.1502	**0.2057**	0.0981	0.0004	0.0053	0.1866
+ SGV	0.4462	0.1149	**0.0939**	**0.0909**	0.0939	**0.1784**	0.0578	0.1129	**0.0303**	**0.0169**	0.2200
+ AGIM	0.4500	0.1748	0.0081	**0.0909**	0.4992	0.1133	0.0421	0.1065	0.0059	0.0043	0.2324
+ MIG	**0.4773**	0.2377	0.0077	0.0642	**0.5413**	0.1315	0.0412	**0.1476**	0.0023	0.0243	**0.2511**
Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Results
	CorLoc										
Baseline	0.7698	0.2894	0.9266	0.6301	0.1300	0.9009	**0.0021**	**0.1696**	0.3788	0.4462	0.4334
+ SGV	0.7698	0.2894	**0.9566**	0.6301	**0.2300**	0.9409	**0.0021**	**0.1696**	**0.5788**	0.4462	0.4519
+ AGIM	**0.8255**	**0.4839**	**0.9566**	**0.7113**	0.0732	0.8695	**0.0021**	0.1571	0.4685	0.4989	0.4512
+ MIG	0.7698	0.4686	0.9539	0.6361	**0.2300**	**0.9507**	**0.0021**	**0.1696**	**0.5788**	**0.5077**	**0.4680**
Methods	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	Results
	CorLoc										
Baseline	0.8808	**0.4939**	0.1565	0.2819	**0.9828**	0.8297	0.2275	0.1034	0.0462	0.0223	0.4334
+ SGV	0.8808	**0.4939**	0.1565	0.2819	**0.9828**	0.8297	0.2275	0.1034	0.0462	0.0223	0.4519
+ AGIM	**0.9030**	0.4530	0.1793	0.3725	0.9776	**0.8645**	0.0850	0.0020	0.0912	0.0483	0.4512
+ MIG	0.8939	0.4212	**0.1978**	**0.3794**	0.9793	0.8065	**0.1377**	**0.1034**	**0.1050**	**0.0694**	**0.4680**

boosted to 23.24% and CorLoc is boosted to 45.12% by only applying AGIM. Besides, it can be seen that collaborating SGV and AGIM further improves the detection performance by 1.87% mAP (25.11% versus 23.24%) and 1.68% CorLoc (46.8% versus 45.12%), respectively. It is mainly because SGV not only facilitates the WSOD model to mine high-quality instances but also encourages the AGIM model to discover all possible same-category instances accurately through generating a more appropriate interclass similarity threshold. It fully demonstrated the effectiveness of each module.

4) Effect of the Number of Clusters in K-Means: In the proposed SGV module, we used the K-means algorithm to divide the proposals into some clusters and define the highest score center as vertexes of the graph. As the number of clusters is a crucial parameter for the K-means algorithm, we analyze how the number of clusters affects the detection performance in Table VI. We can observe that three clusters outperform others. When the number of clusters is smaller than 3, few numbers clusters will introduce many false-positive instances, which hurts the discriminative of the detection network. When we set a larger number of clusters, we will get more top-ranking proposals, which makes the sample number of the SGV model insufficient and the confidence of the voting process reduced, and results in a worse performance than the three clusters.

D. Comparisons With State of the Arts

Here, we provide comprehensive comparisons for each class with previous popular WSOD methods and classical supervised approaches on both the NWPU VHR-10.v2 dataset and the DIOR dataset. We first report our results on the NWPU VHR-10.v2 dataset in Tables II and III. Our method achieves start-of-the-art mAP (55.95%) and the second best CorLoc (70.16%) and significantly boosts the baseline work [59] by a large margin (i.e., mAP + 13.66% and CorLoc + 20.09%). Compared to other previous WSOD methods except for PCIR [22], our method outperforms them by at least 3.84% (55.95% versus 52.11%) in terms of mAP on the test set and 25.10% (70.16% versus 45.06%) in terms of CorLoc on the trainval set. Moreover, we achieve comparable performance with PCIR [22]. Our MIG outperforms PCIR in terms of mAP on the test set (55.95% versus 54.97%) but slightly underperforms PCIR in terms of CorLoc on the trainval set (70.16% versus 71.87%).

We indicate quantitative comparisons with existing popular WSOD methods on the more challenging DIOR dataset in terms of mAP and CorLoc. As clearly reported in Tables IV and V, we achieve the start-of-the-art mAP (25.11%) and CorLoc (46.8%), and notably outperform the baseline work [59] by 6.45% and 3.46%. For most previous WSOD methods, i.e., WSDDN [12], OICR [17], PCL [20], and DCL [23], our method outperforms them by at least 4.92% (25.11% versus 20.19%) in terms of mAP on the test set. Compared with PCIR [22] that achieves the best performance among previous works, our method achieves comparable performance to it (25.11% versus 24.92%), while no extra hyperparameters are introduced. Specifically, the detection performance for "airport" (+49.34%), "bridge" (+8.41%),

TABLE II
PERFORMANCE COMPARISONS (CORLOC) AMONG DIFFERENT METHODS ON THE NWPU VHR-10.V2 TRAINVAL SET

Methods	Airplane	Ship	Storage tank	Baseball Diamond	Tennis court	Basketball court	Ground trackfield	Harbor	Bridge	Vehicle	CorLoc
WSDDN [12]	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	0.8660	0.0741	0.0370	0.1444	0.4001
PCL [20]	0.1176	0.5000	0.1282	0.9865	**0.8451**	0.7736	0.9072	0.0000	0.0926	0.1556	0.4506
MELM [59]	0.8596	0.7742	0.2143	0.9833	0.1071	0.4348	0.9500	0.4000	0.1176	0.1463	0.4987
PCIR [22]	**1.0000**	**0.9306**	0.6410	**0.9932**	0.6479	**0.7925**	0.8969	0.6296	**0.1326**	**0.5222**	**0.7187**
Ours	0.9779	0.9026	**0.8718**	0.9865	0.5493	0.6415	**1.0000**	**0.7407**	0.1296	0.2157	0.7016

TABLE III
PERFORMANCE COMPARISONS (AP AND MAP) AMONG DIFFERENT METHODS ON THE NWPU VHR-10.V2 TEST SET

Methods	Supervision	Airplane	Ship tank	Storage diamond	Baseball court	Tennis court	Basketball trackfield	Ground	Harbor	Bridge	Vehicle	mAP
COPD [67]	Fully Supervised	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 [44]		0.8871	0.7834	0.8633	0.8909	0.4233	0.5685	0.8772	0.6747	0.6231	0.7201	0.7311
RCNN [26]		0.8537	0.8888	0.6278	0.1973	0.9066	0.5823	0.6795	0.7987	0.5422	0.4992	0.6576
Fast RCNN [27]		0.9091	0.9060	0.8929	0.4732	**1.0000**	**0.8585**	0.8486	**0.8822**	**0.8029**	0.6984	0.8271
Faster RCNN [28]		0.9090	0.8630	0.9053	**0.9824**	0.8972	0.6964	**1.0000**	0.8011	0.6149	0.7814	0.8451
RICO [11]		**0.9970**	**0.9080**	**0.9061**	0.9291	0.9029	0.8013	0.9081	0.8029	0.6853	**0.8714**	**0.8712**
WSDDN [12]	Weakly Supervised	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
PCL [20]		0.2600	0.6376	0.0250	0.8980	0.6445	0.7607	0.7794	0.0000	0.0130	0.1567	0.3941
MELM [59]		0.8086	0.6930	0.1048	0.9017	0.1284	0.2014	0.9917	0.1710	0.1417	0.0868	0.4229
DCL [23]		0.7270	0.7425	0.3705	0.8264	0.3688	0.4227	0.8395	0.3957	**0.1682**	0.3500	0.5211
PCIR [22]		**0.9078**	**0.7881**	0.3640	0.9080	0.2264	**0.5216**	0.8851	**0.4236**	0.1174	**0.3549**	0.5497
Ours		0.8869	0.7161	**0.7517**	**0.9419**	**0.3745**	0.4768	**1.0000**	0.2727	0.0833	0.0906	**0.5595**

TABLE IV
PERFORMANCE COMPARISONS (CORLOC) AMONG DIFFERENT METHODS ON THE DIOR TRAINVAL SET

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	CorLoc
WSDDN [12]	0.0572	**0.5988**	0.9424	0.5594	0.0492	0.2340	**0.0103**	0.0679	0.4452	0.1275	0.3244
OICR [17]	0.1598	0.5145	0.9477	0.5579	0.0355	0.2389	0.0000	0.0482	0.5668	0.2242	0.3477
PCL [20]	0.6114	0.4686	**0.9539**	**0.6361**	0.0732	**0.9507**	0.0021	0.0571	0.0514	**0.5077**	0.4152
MELM [59]	**0.7698**	0.2894	0.9266	0.6301	0.1300	0.9009	0.0021	**0.1696**	0.3788	0.4462	0.4334
Ours	**0.7698**	0.4686	**0.9539**	**0.6361**	**0.2300**	**0.9507**	0.0021	**0.1696**	**0.5788**	**0.5077**	**0.4680**
Methods	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	CorLoc
WSDDN [12]	0.8990	0.0545	0.1000	0.2296	**0.9854**	0.7961	0.1506	0.0345	0.1156	0.0322	0.3244
OICR [17]	**0.9141**	0.1818	0.1870	0.3180	0.9828	0.8129	0.0745	0.0122	**0.1583**	0.0198	0.3477
PCL [20]	0.8939	0.4212	**0.1978**	**0.3794**	0.9793	0.8065	0.1377	0.0020	0.1050	**0.0694**	0.4152
MELM [59]	0.8808	**0.4939**	0.1565	0.2819	0.9828	**0.8297**	**0.2275**	**0.1034**	0.0462	0.0223	0.4334
Ours	0.8939	0.4212	**0.1978**	**0.3794**	0.9793	0.8065	0.1377	**0.1034**	0.1050	**0.0694**	**0.4680**

"chimney" (+4.91%) "ground track field" (+6.68%), "golf field" (+42.13%), and "stadium" (+45.55%) is significantly boosted since the OICR-based instance mining strategy is limited by the quality of the initial object proposal. If there are no reasonable proposals for model initialization, this refinement strategy cannot correctly discover the whole object. The large improvement for the above large object categories fully exhibits the general effectiveness of the proposed method for WSOD. Besides, comparisons with previous WSOD approaches in terms of CorLoc are indicated in Table IV. We can clearly see the similar phenomenon that our method surpasses other methods by a large margin.

To fully testify the effectiveness of the proposed method, we also provide comparisons with some classical full-supervised methods in RSIs, including the collection of the part detector (COPD) [67], the transferred CNN model from AlexNet [7], the RICNN [44], RCNN [26], Fast-RCNN [27], and Faster-RCNN [28]. As exhibited in Tables III and V, we achieve comparable and even exceeded performance than some full-supervised learning methods.

Although a large improvement has been achieved for the average performance, our approach fails to detect individual classes including the dam, overpass, tennis court, and wind mill. There are two main reasons for this result: 1) WSOD

TABLE V
PERFORMANCE COMPARISONS (AP AND MAP) AMONG DIFFERENT METHODS ON THE DIOR TEST SET

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	mAP
Fast-RCNN [27]	0.4417	**0.6679**	**0.6696**	0.6049	0.1556	**0.7228**	**0.5195**	**0.6587**	0.4476	**0.7211**	0.4998
Faster-RCNN [28]	**0.5028**	0.6260	0.6604	**0.8088**	**0.2880**	0.6817	0.4726	0.5851	**0.4806**	0.6044	**0.5548**
WSDDN [12]	0.0906	0.3968	0.3781	0.2016	0.0025	0.1218	0.0057	0.0065	0.1188	0.0490	0.1326
OICR [17]	0.0870	0.2826	0.4405	0.1822	0.0130	0.2015	0.0009	0.0065	**0.2989**	0.1380	0.1650
PCL [20]	0.2152	0.3519	0.5980	0.2349	0.0295	0.4371	0.0012	0.0090	0.0149	0.0288	0.1819
MELM [59]	0.2814	0.0323	0.6251	**0.2872**	0.0006	0.6251	0.0021	**0.1309**	0.2839	0.1515	0.1866
DCL [23]	0.2089	0.2270	0.5421	0.1150	0.0603	0.6101	0.0009	0.0107	0.3101	0.3087	0.2019
PCIR [22]	**0.3037**	0.3606	0.5422	0.2660	**0.0909**	0.5859	0.0022	0.0965	0.3618	0.3259	0.2492
Ours	0.2220	**0.5257**	**0.6276**	0.2578	0.0847	**0.6742**	**0.0066**	0.0885	0.2871	**0.5728**	**0.2511**

Methods	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	mAP
Fast-RCNN [27]	0.6293	**0.4618**	0.3803	0.3213	**0.7098**	0.3504	0.5827	0.3791	0.1920	0.3810	0.4998
Faster-RCNN [28]	**0.6700**	0.4386	**0.4687**	**0.5848**	0.5237	**0.4235**	**0.7952**	**0.4802**	**0.3477**	**0.6544**	**0.5548**
WSDDN [12]	0.4235	0.0466	0.0106	0.0070	0.6303	0.0395	0.0606	0.0051	0.0455	0.0114	0.1326
OICR [17]	0.5739	0.1066	0.1106	0.0909	0.5929	0.0710	0.0068	0.0014	**0.0909**	0.0041	0.1650
PCL [20]	0.5636	0.1676	0.1105	0.0909	0.5762	0.0909	0.0247	0.0012	0.0455	0.0455	0.1819
MELM [59]	0.4105	**0.2612**	0.0043	0.0909	0.0858	**0.1502**	**0.2057**	0.0981	0.0004	0.0053	0.1866
DCL [23]	0.5645	0.0505	0.0265	0.0909	0.6365	0.0909	0.1036	0.0002	0.0727	0.0079	0.2019
PCIR [22]	**0.5851**	0.0860	**0.2163**	**0.1209**	**0.6428**	0.0909	0.1362	0.0030	**0.0909**	**0.0752**	0.2492
Ours	0.4773	0.2377	0.0077	0.0642	0.5413	0.1315	0.0412	**0.1476**	0.0023	0.0243	**0.2511**

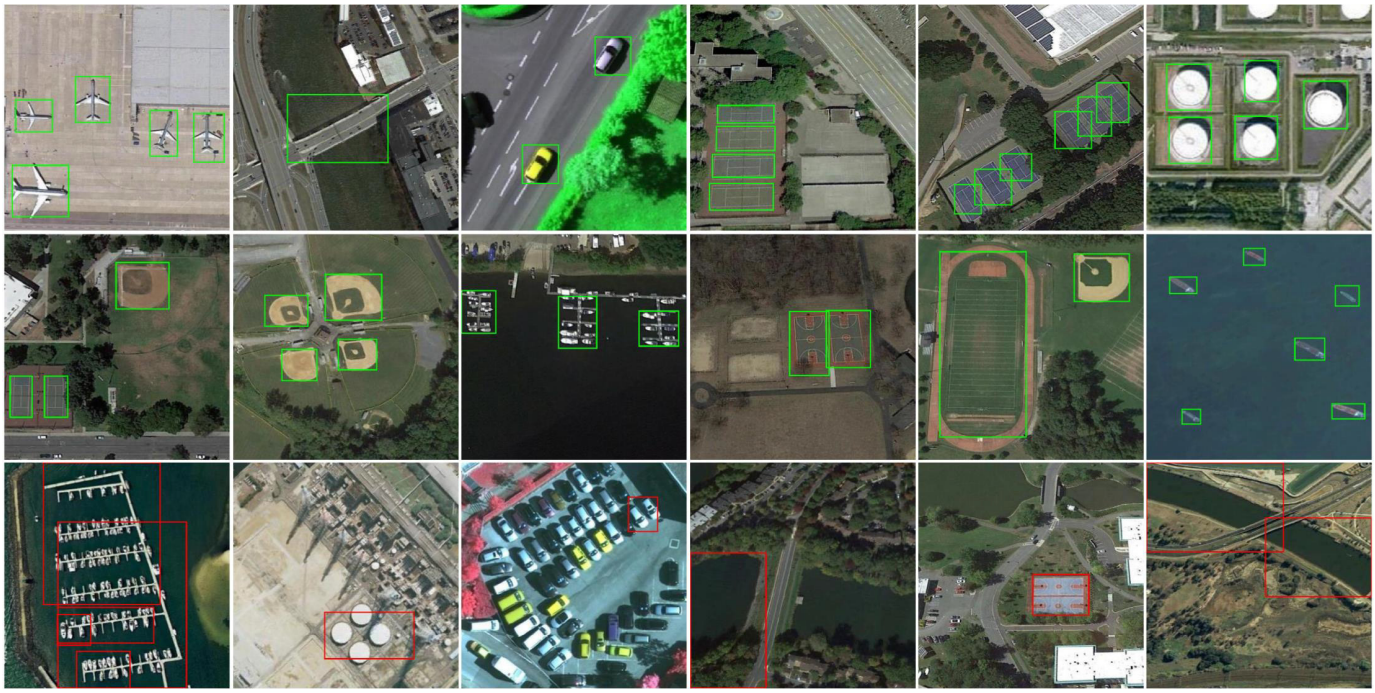

Fig. 3. Example results on the NWPU VHR-10.V2 test. The first two rows indicate success cases with a green color rectangle, and the last row denotes a few missed objects and false positives with a red color rectangle.

model inclines to discover more salient objects when object category always coexisting with a specific background in an image, thereby mistakenly identifying the special background (i.e., rivers, roads, and reservoirs) as the object (bridges, overpass, and dams) and 2) multiple instances belonging to the same category are grouped into a single bounding box. They often appear in adjacent locations.

We finally provide qualitative results by visualizing some success and failure detection examples on the NWPU VHR-10.v2 dataset and the DIOR dataset. As shown in the first two rows of Figs. 3 and 4, the predicted bounding boxes can accurately and tightly cover all objects. However, as shown in the last row of Figs. 3 and 4, our method detects multiple adjacent objects as ones and mistakenly identifies salient context background as objects. An ideal solution is yet needed as there is still room for improvement. Moreover, we provide a qualitative comparison between the baseline and the proposed method in Fig. 5. We can observe that our

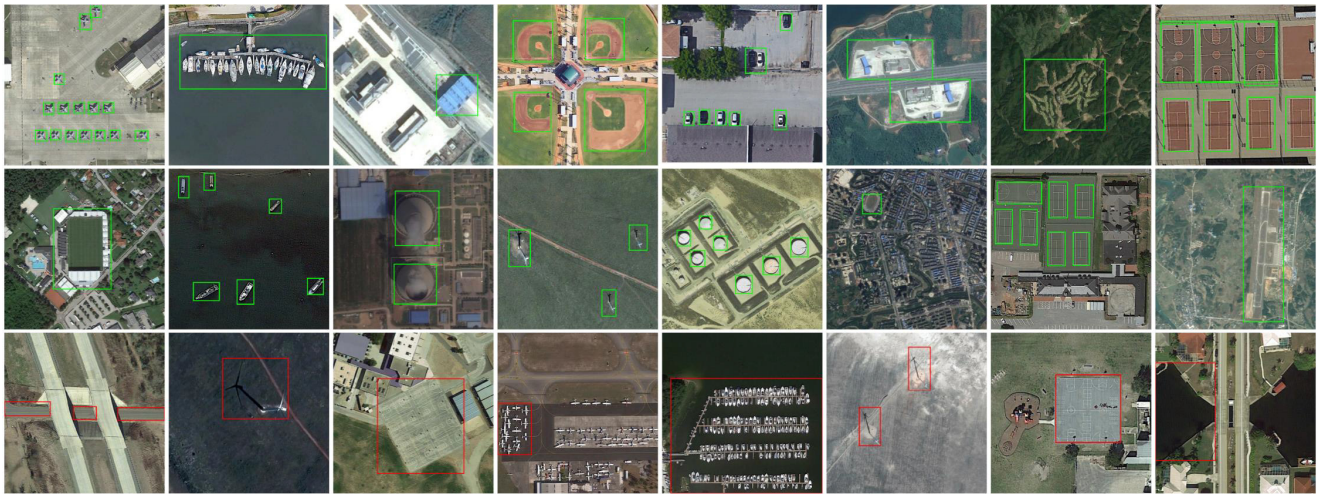

Fig. 4. Example results on the DIOR test. The first two rows indicate success cases with a green color rectangle, and the last row denotes a few missed objects and false positives with a red color rectangle.

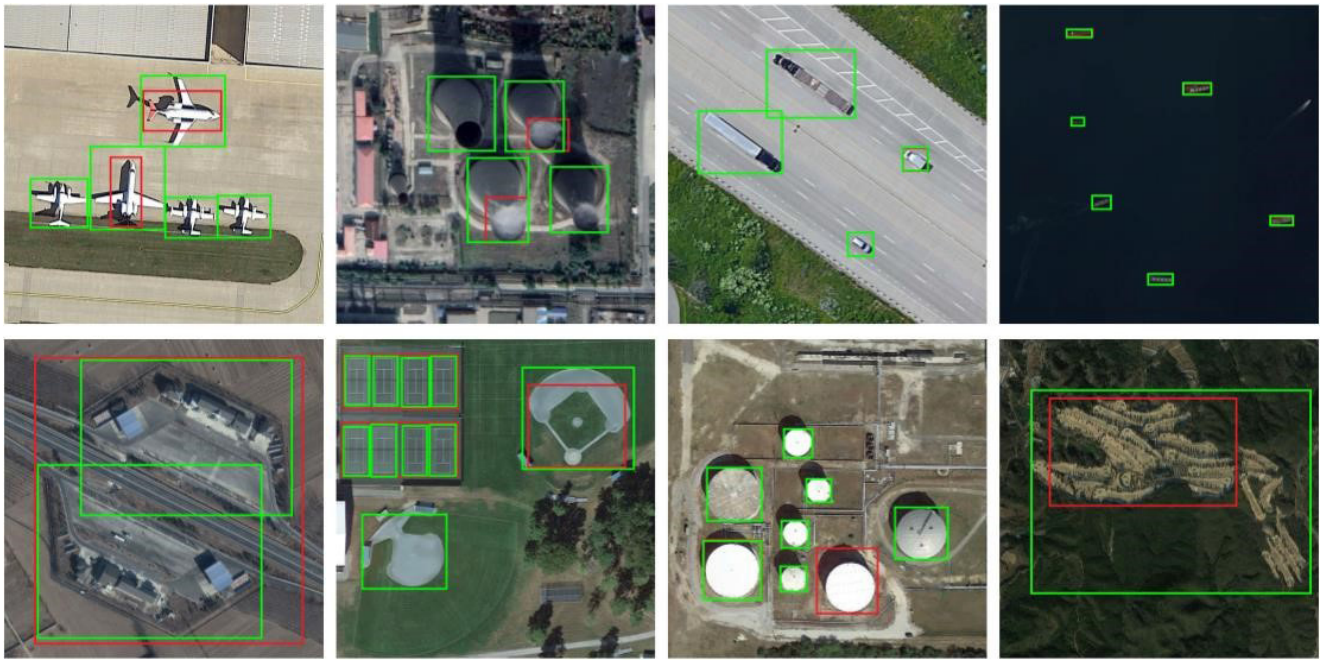

Fig. 5. Qualitative comparisons with a baseline work. The red color rectangle denotes the results of the baseline work, and the green color rectangle denotes the results of the proposed method.

TABLE VI

ANALYSIS OF CLUSTER NUMBER ON THE DIOR DATASET. WE REPORT THE MAP AND CORLOC OF THE SGV MODULE

Cluster number	2	3	4	5
mAP(%)	20.8	22.0	21.6	19.9
CorLoc(%)	43.8	45.2	44.6	43.2

TABLE VII

RUNTIME ANALYSIS ON THE DIOR DATASET

Method	**mAP (%)**	**speed (FPS)**
Baseline	18.7	3.2
+ SGV	22.0	2.7
+ AGIM	23.2	2.9
+ MIG	25.1	2.6

method can achieve better performance and mine all possible same-category instances in an image and further demonstrate the effectiveness of the proposed method.

E. Runtime Analysis

As shown in Table VII, we conduct experiments to analyze how each module in the proposed MIG model affects the efficiency of the proposed method. To pursue all possible high-quality same-class instances, we first propose a novel spatial graph-based vote algorithm to pursue high-quality objects by collecting spatial votes from top-ranking proposals with highly spatial overlap with it. Then, we further construct an appearance graph-based instance mine strategy,

which aims to discover all possible instances according to the apparent similarities. Thus, compared with our baseline work, our method introduces additional computation costs. As shown in Table VII, compared with our baseline work, the computational efficiency dropped from 3.23 to 2.74 frames per second (FPS) by integrating the SGV into the baseline work. The additional calculations are mainly introduced by the cluster operations. Besides, when we integrate AGIM into the baseline, the computational efficiency dropped from 3.23 to 2.91 FPS, which is caused by an apparent similarity calculation. Although the baseline work is slightly faster than our method (3.23 versus 2.56 FPS), but its accuracy is reduced by 6.4%.

V. Conclusion

In this article, we propose a novel MIG learning framework to train more robust refined instance classifiers for WSOD in RSIs. Specifically, we first build a spatial graph according to their spatial correlation and treat the cluster center with the most spatial votes as pseudoinstance-level labels. Then, we introduce an appearance graph-based instance mine strategy to mine more possible instances of the same class via propagating the label information from the obtained high-quality objects to other graph nodes according to the apparent similarity. Combining SGV with AGIM formulates novel multiple instance mining strategies to pursue all possible object instances appearing in an image. The algorithm is independent of any extra hyperparameters and annotations. Experiments on two public remote sensing datasets show substantial and consistent improvements by our method. We also find that our approach is a failure to coexisting and adjacent objects. In the future, we will concentrate on these challenges.

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] Y. Bu *et al.*, "Hyperspectral and multispectral image fusion via graph laplacian-guided coupled tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 648–662, Jan. 2021.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[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] Z. Chen, S. Huang, and D. Tao, "Context refinement for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 71–86.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3052–3062.

[26] 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.

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

[28] 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.

[29] Z. Huang, W. Li, X.-G. Xia, X. Wu, Z. Cai, and R. Tao, "A novel nonlocal-aware pyramid and multiscale multitask refinement detector for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Feb. 26, 2021, doi: 10.1109/TGRS.2021.3059450.

[30] L. Li, X. Yao, G. Cheng, M. Xu, J. Han, and J. Han, "Solo-to-collaborative dual-attention network for one-shot object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 2, 2021, doi: 10.1109/TGRS.2021.3091003.

[31] Z. Huang, W. Li, X.-G. Xia, H. Wang, F. Jie, and R. Tao, "LO-Det: Lightweight oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Mar. 3, 2021, doi: 10.1109/TGRS.2021.3067470.

[32] S. Tian, L. Kang, X. Xing, J. Tian, C. Fan, and Y. Zhang, "A relation-augmented embedded graph attention network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, early access, May 18, 2021, doi: 10.1109/TGRS.2021.3073269.

[33] M. Zand, A. Etemad, and M. Greenspan, "Oriented bounding boxes for small and freely rotated objects," *IEEE Trans. Geosci. Remote Sens.*, early access, May 5, 2021, doi: 10.1109/TGRS.2021.3076050.
[34] G. Guo, Z. Liu, S. Zhao, L. Guo, and T. Liu, "Eliminating indefiniteness of clinical spectrum for better screening COVID-19," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1347–1357, May 2021.
[35] L. Yang, H. Peng, D. Zhang, J. Fu, and J. Han, "Revisiting anchor mechanisms for temporal action localization," *IEEE Trans. Image Process.*, vol. 29, pp. 8535–8548, 2020.
[36] 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.
[37] 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)*, Jun. 2002, pp. 524–531.
[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
[39] 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.
[40] 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. Photogram. Remote Sens.*, vol. 89, pp. 37–48, Mar. 2014.
[41] 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.*, Jun. 2015, pp. 1173–1181.
[42] 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.
[43] 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.
[44] 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.
[45] 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.
[46] 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.
[47] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang, "Group-wise semantic mining for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1984–1992.
[48] G. Guo, J. Han, F. Wan, and D. Zhang, "Strengthen learning tolerance for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 7403–7412.
[49] T. Zhao, J. Han, L. Yang, B. Wang, and D. Zhang, "SODA: Weakly supervised temporal action localization based on astute background response and self-distillation learning," *Int. J. Comput. Vis.*, vol. 109, pp. 2474–2498, May 2021.
[50] Y. Shen *et al.*, "Enabling deep residual networks for weakly supervised object detection," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 118–136.
[51] L. Song, J. Liu, M. Sun, and X. Shang, "Weakly supervised group mask network for object detection," *Int. J. Comput. Vis.*, vol. 129, no. 3, pp. 681–702, Mar. 2021.
[52] Y. Yin, J. Deng, W. Zhou, and H. Li, "Instance mining with class feature banks for weakly supervised object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3190–3198.
[53] 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.
[54] D. Li, J. Huang, Y. Li, S. Wang, and M. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 3512–3520.
[55] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep self-taught learning for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4294–4302.
[56] Q. Ye, F. Wan, C. Liu, Q. Huang, and X. Ji, "Continuation multiple instance learning for weakly and fully supervised object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 16, 2021, doi: 10.1109/TNNLS.2021.3070801.
[57] Q. Meng, W. Wang, T. Zhou, J. Shen, L. Van Gool, and D. Dai, "Weakly supervised 3D object detection from lidar point cloud," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 515–531.
[58] Y. Xu, C. Zhou, X. Yu, B. Xiao, and Y. Yang, "Pyramidal multiple instance detection network with mask guided self-correction for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3029–3040, 2021.
[59] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1297–1306.
[60] 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.
[61] 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.
[62] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Learning rotation-invariant local binary descriptor," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3636–3651, Aug. 2017.
[63] R. Jiang, S. Mei, M. Ma, and S. Zhang, "Rotation-invariant feature learning in VHR optical remote sensing images via nested Siamese structure with double center loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3326–3337, Apr. 2021.
[64] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," 2020, *arXiv:2008.09397*.
[65] 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, Apr. 2013.
[66] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.
[67] 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.
[68] S. Karen and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[69] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

Binglu Wang (Member, IEEE) received the M.S. degree in robotics from University West, Trollhättan, Sweden, in 2016, and the Ph.D. degree in control science and engineering from the School of Automation, Northwestern Polytechnic University, Xi'an, China, in 2021.

His research interests include computer vision, robotics science, and deep learning.

Yongqiang Zhao (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in control science and engineering from Northwestern Polytechnic University, Xi'an, China, in 1998, 2001, and 2004, respectively.

From 2007 to 2009, he was a Post-Doctoral Researcher with McMaster University, Hamilton, ON, Canada, and Temple University, Philadelphia, PA, USA. He is currently a Professor with Northwestern Polytechnical University. His research interests include polarization vision, hyperspectral imaging, and pattern recognition.

Xuelong Li (Fellow, IEEE) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.