Point-Based Weakly Supervised Learning for Object Detection in High Spatial Resolution Remote Sensing Images

Youyou Li, Binbin He, *Member, IEEE*, Farid Melgani, *Fellow, IEEE*, and Teng Long

***Abstract*—Object detection is challenging in high spatial resolution (HSR) remote sensing images that have a complex background and irregular object locations. To minimize manual annotation cost in supervised learning methods and achieve advanced detection performance, we proposed a point-based weakly supervised learning method to address the object detection challenge in HSR remote sensing images. In the study, point labels are introduced to guide candidate bounding box mining and generate pseudobounding boxes for objects. Then, pseudobounding boxes are applied to train the detection model. A progressive candidate bounding box mining strategy is proposed to refine object detection. Experiments are conducted on a comprehensive HSR dataset which contains four categories. Results indicate the proposed method achieves competitive performance compared to YOLOv5 which is trained on manual bounding box annotations. In comparison to the state-of-the-art weakly supervised learning method, our method outperforms WSDDN method with 0.62 mean average precision score.**

***Index Terms*—High spatial resolution (HSR), object detection, point-based supervision, remote sensing, weakly supervised learning.**

I. Introduction

OWING to the rapid development of remote sensing platforms and the reduction of data collection costs, a growing number of high spatial resolution (HSR) remote sensing images are publicly available [1], [2]. Interpreting HSR remote sensing images is crucial since much more detailed information can be obtained from HSR remote sensing images compared to that with relatively lower spatial resolution. Object detection is a part of the most useful techniques for understanding HSR remote sensing images. In the last decades, due to the rarity of HSR remote sensing images, researches mainly focused on detecting big categories such as crops, forests, and urban landscapes, etc. [3]–[5]. More recently, since access to HSR remote sensing image is easier, and detecting objects is meaningful for practical applications such as urban planning, transportation management, energy assessment, etc., detecting objects such as vehicles, ships, and storage tanks has aroused great interest in the field of remote sensing [6], [7].

Though HSR remote sensing images offer possibilities for detecting objects, the characteristic of HSR brings new challenges to identify and localize objects. With the increase in spatial resolution, remote sensing images can present more details of the land surface, so the scenes of HSR remote sensing images become complex and heterogeneous [8]. Heterogeneity of pixels adds difficulties to identify the edges of objects and increases intraclass variety within homogeneous classes. As HSR remote sensing images may be obtained under different acquisition conditions, the size, color, and view of an object are easily influenced and vary. Another significant influence is rotation variations, which are notoriously difficult to efficiently deal with using existing techniques [9]. Moreover, shadows from trees and buildings in HSR remote sensing images often cover lots of objects and make their appearance blur and vague. This makes the covered object difficult to locate and detect [10]. In order to address the above problems, researchers have proposed fully supervised learning algorithms for object detection in HSR remote sensing images.

Fully supervised learning methods are widely implemented to detect objects in HSR remote sensing images. By leveraging the development of graphic process units and the availability of massive data, deep convolutional neural network (DCNN) experimentally turned out to be effective in dealing with recognition challenges. A large number of methods based on DCNN are proposed to carry out object detection in HSR remote sensing images. For example, taking full advantage from fully convolutional network, Zhong *et al.* addressed the dilemma between translation-invariance in the classification stage and translation-variance in the object detection stage in HSR images [11]. Gong *et al.* proposed a context-aware convolutional neural network model to enhance the representativeness of features for object detection in very high resolution remote sensing images [12]. Though fully supervised DCNN achieved satisfying performance on object detection task, the limitation of fully

Manuscript received February 17, 2021; revised March 31, 2021; accepted April 24, 2021. Date of publication April 27, 2021; date of current version June 8, 2021. This work was supported in part by the Department of Science, and Technology of Sichuan Province under Grant 2020YFS0058, and in part by the Special Project of Local Science, and Technology Development Guided by the Central Government in 2020 under Grant 2020ZYD094. *(Corresponding authors: Binbin He; Farid Melgani)*

Youyou Li and Binbin He are with the School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: li_youyou@std.uestc.edu.cn; binbinhe@uestc.edu.cn).

Farid Melgani is with the Department of Information Engineering and Computer Science, University of Trento, I-38123 Trento, Italy (e-mail: melgani@disi.unitn.it).

Teng Long is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: uestc.longteng@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2021.3076072

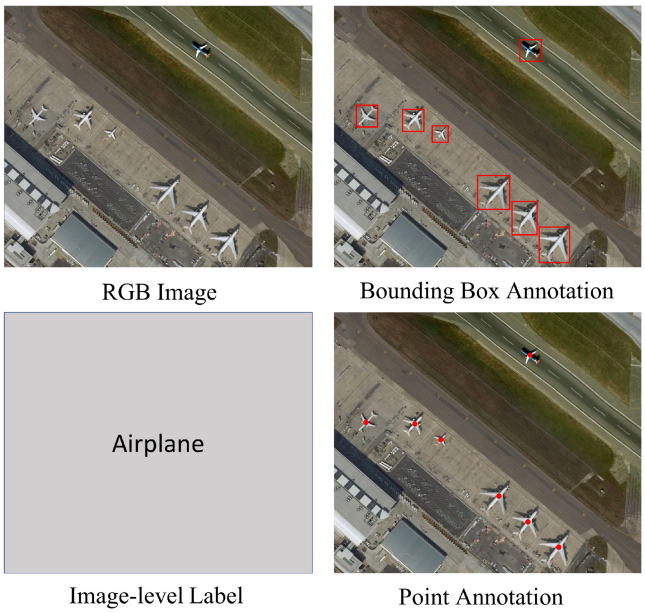

Fig. 1. Illustration of bounding box annotation, image-level label, and point annotation.

supervised DCNNs is also evident that fully supervised DCNNs need lots of richly annotated labels which require intensive and time-consuming manual annotation. Especially, drawing bounding boxes for object detection training data is both subjective and expensive. Therefore, an alternative training pattern is therefore suggested sidestepping the limitation posed by human labeling.

In order to decrease the amount of human intervention needed for training models, weakly supervised learning (WSL) is introduced to computer vision systems [13]. WSL means the method is trained using examples that are only partially annotated or labeled [14]. One main type of partial labels used to annotate instances is image-level label which means that only the image as a whole is annotated with meaning. An example of image-level labels is presented in Fig. 1. The image-level label of an image can only tell us the image contains specific objects, while the locations of these objects, the number of these objects contained in the image, and the size and boundary of each object are unknown. Though it is challenging to supervise DCNNs using only image-level labels, many researchers actively explored the possibility using image-level labels for training detection models. Since one image may contain many instances of a specific object but the image is only labeled positive, retrieving the location, size, and boundary of individual instances in the image is necessary for object detection. Multiple instance learning (MIL) is a dominant method to predict whether each instance proposal of an image contains the instance or not [15]. Andrews *et al.* have proposed an SVM-based approach for MIL, which is regarded as a standard MIL form in the following studies [16]. To detect objects, many studies follow the standard MIL in their WSL systems [17]–[20], which usually iteratively implement two-steps operations. The first stage is to generate bounding box proposals from unsupervised learning methods such as selective search [21] and to learn the top-scoring proposals for each image under MIL constraints; the second stage is to train detection models and update proposals taking advantage from detection results.

Although encouraging results are reported, there are two main problems in image-level label-based WSL methods for object detection in remote sensing images. First, in the aforementioned methods, most of the WSL methods tend to select one proposal for the corresponding object category for each image. Distinct from natural images, one remote sensing image often contains multiple same-class instances [22]. Only choosing one proposal of each class within an image is insufficient to train detection models. Moreover, WSL methods can hardly mine high-quality proposals with image-level labels. However, top-scoring instance proposals generated from the aforementioned methods only cover a part of objects or even cannot locate objects. Thus, image-level labels are deficient for training WSL detection models. Second, the aforementioned methods do not constitute an alternative to fully supervised learning methods, since the performance that WSL methods based on image-level labels can achieve is much lower than fully supervised learning methods. Also, accuracy that WSL methods with image-level labels can obtain may not be sufficient for practical applications.

In order to prevent cumbersome bounding box annotation process and to overcome the challenges raised by the aforementioned WSL methods, we propose a point-based weakly supervised learning method for detecting objects in HSR remote sensing images. The concept of point annotation (showed in Fig. 1)-based WSL methods are first introduced by Pascal *et al.* who proposed a pointly supervised method for localizing actions in videos, which obtained similar action localization performance to bounding box-based supervised learning methods [23]. There are several essential differences between action localization in videos and object detection in HSR remote sensing images. First, usually a video only contains one action, while an HSR remote sensing image often contains multiple instances related to one or more classes. For example, a video is about skateboarding; an HSR remote sensing image contains numerous cars and storage tanks. Second, temporal information is important for action localization in videos, while object detection in HSR remote sensing images needs to focus on spatial information. Therefore, it is necessary to specifically design a point-based weakly supervised learning method for object detection in HSR remote sensing images.

In this study, the proposed point-based weakly supervised learning method aims to solve multiclass object detection tasks in HSR remote sensing images. The method mainly includes two steps: Progressively proposal mining and training detection models. In proposal mining step, three normalized measurement scores are proposed to specifically measure proposals generated from HSR remote sensing images. Pseudobounding box labels are then generated from sorting proposals. In the training step, a modified CIoU loss is introduced to balance the side effect brought by pseudolabels, and detection models are trained on the YOLOv5 network [24], [25]. We compare the proposed point-based weakly supervised learning method to two state-of-the-art methods which are a fully supervised learning method and an

image-level label-based WSL method [26] to demonstrate the effectiveness of the proposed method.

We summarize the main contributions of this article as follows.

1) Propose a new point-based supervised learning method to detect multiclass objects in HSR remote sensing images.
2) Propose three measurement methods, respectively, to assess the distance between the center of proposals to point labels, the side length of proposals, and the comprehensive performance of proposals for HSR remote sensing images.
3) Propose a weighted CIoU loss to neutralize the side effect introduced by pseudolabels.

The remainder of this article is organized as follows. Section II presents the principles of proposed point-based weakly supervised learning method. Section III describes datasets used in this study, the experiment configuration and the experimental results achieved on four categories. Section IV concludes this article.

II. Proposed Method

A. Proposal Measurement

Before depicting our method, following notations used in the article are introduced first. For an image I, $I^{(w)}$ and $I^{(h)}$, respectively, represents the width and height of I. $\boldsymbol{P} = \{\boldsymbol{BB}\}^n$ represents proposals of I, where $1 \leq n$, n is the number of proposals of the image. $(c_i^{(x)}, c_i^{(y)})$ denotes the point annotation related to the ith instance in I. $\boldsymbol{P_i} = \{\boldsymbol{BB}_i\}^m$ represents proposals of I containing ith point label, where $1 \leq m \leq n$, m is the number of proposals which contain the ith point annotation. C_i denotes the category of the ith point label. $(BB_{ij}{}^{(x)}, BB_{ij}{}^{(y)})$ represents the left-top point of the jth bounding box of $\boldsymbol{P_i}$. $(BB_{ij}{}^{(c_x)}, BB_{ij}{}^{(c_y)})$ represents the center point of the jth bounding box of $\boldsymbol{P_i}$. $(BB_{ij}{}^{(w)}, BB_{ij}{}^{(h)})$ represents the width and height of the jth bounding box of $\boldsymbol{P_i}$. D_{ij} denotes the distance measure of the jth bounding box of $\boldsymbol{P_i}$. S_{ij} denotes the size measure of the jth bounding box of $\boldsymbol{P_i}$. O_{ij} represents the overall measure of the jth bounding box of $\boldsymbol{P_i}$. $P_{gt}^{(i)}$ denotes the pseudobounding box corresponding to the ith point labeled with C_i category.

The proposal measurement is comprised of a normalized overall measure (NOM), a normalized distance measure, and a normalized size measure. The distance measure describes how near the center point of proposals to point annotations. It is clear the smaller the distance the better the proposals. The size measure describes the ratio between the width and height of proposals and the width and height of HSR remote sensing images. Since proposals with larger size have a high possibility of containing multiple objects, the size measure follows the hypothesis that the shorter the width and height of the bounding box, the more homogeneous pixels it contains. By combining distance and size measure, NOM is designed to comprehensively estimate the quality of proposals.

1) Normalized Distance Measure: For the jth proposal of $\boldsymbol{P_i}$, the distance term D_{ij} is defined as the ratio of the distance between the center point of a bounding box and the corresponding point label. To normalize it, in (1), the distance is divided by half the diagonal length of the bounding box, because the distance between points inside a rectangle and the center point of the rectangle ranges from 0 to half the diagonal length of the rectangle. Therefore, the normalized distance ratio ranges 0 to 1. The smaller the ratio, the closer the distance between the two points

$$D_{ij} = \frac{\left\| (c_i^{(x)}, c_i^{(y)}) - (BB_{ij}{}^{(c_x)}, BB_{ij}{}^{(c_y)}) \right\|_2}{\left\| (BB_{ij}{}^{(x)}, BB_{ij}{}^{(y)}) - (BB_{ij}{}^{(c_x)}, BB_{ij}{}^{(c_y)}) \right\|_2}. \quad (1)$$

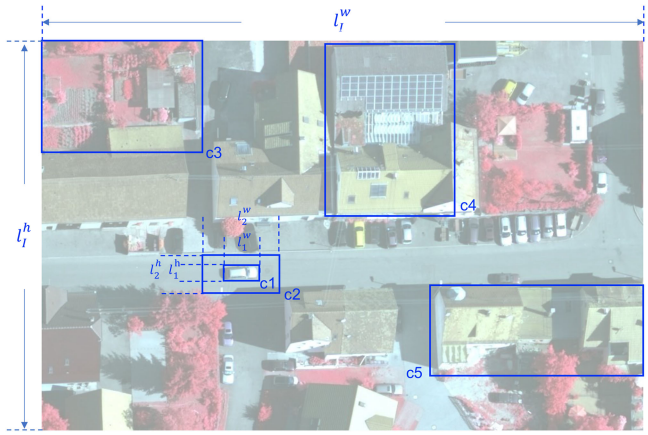

Fig. 2. Example of candidate bounding boxes and its corresponding imagery.

2) Normalized Size Measure: As mentioned above, proposals with large size tend to contain multiple instances and objects instead of a single instance or object. Pascal *et al.* proposed an area ratio to measure the relative area of candidate boxes to that of images in natural images [23]. As shown in Fig. 2, $1, c_1, c_2, \ldots, c_n$ represent candidate bounding boxes of the imagery I. l_i^w denotes the length of the width of the ith candidate bounding box, and l_i^w denotes the length of the height of the ith candidate bounding box. l_I^w and l_I^h, respectively, represents the length of the width and height of the imagery I. The equation of area ratio of candidate bounding boxes and the imagery is shown in (2)

$$\frac{l_i^w \times l_i^h}{l_I^w \times l_I^h}. \quad (2)$$

Since in remote sensing images, the area of an object could be much smaller than that of the image, the area ratio of the bounding box c_1 and the image I can be similar to that of the bounding box c_2 and the image I. This means the area ratios can hardly differentiate the size of bounding box c_1 from bounding box c_2. Moreover, area ratio tends to yield unbalanced area score distribution since many area scores of objects of interest are very small in HSR remote sensing images. So the area ratio would be insensitive to differentiate candidate box size. Therefore, we proposed the side length ratio of bounding boxes and their corresponding imagery to better distinguish candidate bounding boxes by their side lengths. From the (3), the longest side of the

width and height of the candidate bounding box divided by half of the side length of the corresponding remote sensing image is used as the size score. Since the denominator of the side length ratio is $l_I^{\text{w or h}}$ which is much smaller than the denominator of the area ratio which is $l_I^w \times l_I^h$, the lengths of box sides of candidate bounding boxes would have a bigger influence on the side ratio than on the area ratio. Thus, the side ratio is much more sensitive to the difference of size of candidate bounding boxes. Therefore, the side ratio is considered as the size measure for candidate bounding boxes

$$S_{ij} = \text{Max}\left(\frac{2BB_{ij}{}^{(w)}}{I^{(w)}}, \frac{2BB_{ij}{}^{(h)}}{I^{(h)}}\right). \tag{3}$$

The normalized size term S_{ij} for the jth bounding box of $\boldsymbol{P_i}$ is evaluated by the maximum value of the ratio of the width of candidate box to half of the image width and the ratio of the height of the candidate box to half of the image height, and it ranges from 0 to 1. The closer the normalized size value is to zero, the more homogeneous the pixels in the candidate box.

3) Normalized Overall Measure: For comprehensively assessing candidate boxes, the normalized distance and size term are fused as NOM score. As (4) shows, the NOM score is represented by the geometric mean of $1 - D_{ij}$ and $1 - S_{ij}$ because the geometric mean can effectively eliminate the influence of extreme values. As the distance between proposal center and point label decreases, $1 - D_{ij}$ increases. As the value of side length ratio decrease, $1 - S_{ij}$ increases. Theoretically, when NOM value is bigger, the distance between the proposal center and the point label is closer and the value of the side length ratio is smaller

$$O_{ij} = \sqrt{(1 - D_{ij}) \times (1 - S_{ij})}. \tag{4}$$

B. Self-Supervised Learning

1) Pseudobounding Box Generation: Regarding proposals sorted by NOM, candidate boxes with the highest score have two properties. First, the center point of the candidate box is relatively closer to the label point; second, the size of the candidate box is small, which ensures that the pixels inside the candidate box are homogeneous. However, the candidate box with the highest score cannot be guaranteed to have the width and height which are closest to the real object. Therefore, the width and height of the pseudolabel are temporarily represented by the average of the width and height of top r proposals, and the width and height of the subsequent pseudolabel will be updated through continuous progressive mining. The hyperparameter of ranking ratio r of candidate boxes is proposed, which is the boundary of the selection of candidate boxes. The initial hyperparameter r is randomly set between (0,1), that is, the candidate box with the first r score is adopted to generate the initial pseudoground truth box. The subsequent hyperparameter r will be updated according to later iterations. In (5), the ith pseudobounding box is generated from top r candidate boxes.

$$P_{gt}^{(i)}(x, y, w, h) = \frac{1}{rm} \times \sum_{j=1}^{rm} BB_j^{(i)}(x, y, w, h). \tag{5}$$

Algorithm 1: Progressive Proposal Mining.

Data: Training images $\boldsymbol{It}$ with corresponding point labels $\boldsymbol{Pt_{point}}$. Validation images $\boldsymbol{Iv}$ with corresponding point and box labels $\boldsymbol{Pv_{point}}$ and $\boldsymbol{Pv_{box}}$.

Result: The learned detection model M.

```
Initialization: r, step, mAP_val0 = 0, ValDifference = 0;
Calculating overall score for each candidate box by Eq. (3);
Generating initial pseudo labels Pseudo_gt for training images by Eq. (4);
while ValDifference ≥ 0 do
    Training detection model M using the pseudo labels generated from this iteration;
    Obtain new training weight W;
    Calculating the mAP_val1 of the inference results on validation images;
    ValDifference = mAP_val1 − mAP_val0;
    mAP_val0 = mAP_val1;
    if ValDifference ≥ 0 then
        Updating r: r = r − step;
        Generating new pseudo labels Pseudo_gt for training images by the updated r;
    end
end
```

2) Weighted Loss: In YOLO, the overall loss consists of a regression and two classification loss terms, in this study, we mainly focus on the regression loss since the width and height of pseudoboxes are inaccurate. The original CIoU regression loss is shown in (6)

$$L_{\text{CIOU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha \times v \tag{6}$$

where IoU denotes the intersection over union of predicted box and ground truth box. $\frac{\rho^2(b,b^{gt})}{c^2}$ represents the center distance evaluation term, where $\rho^2(b, b^{gt})$ denotes the distance between the center of predicted box and the ground truth box, and c represents the diagonal distance of the smallest closed area that can contain both the prediction box and the ground truth box. The last term $\alpha \times v$ is used to measure the similarity of the aspect ratio, and α is a weight value.

To balance the effect brought by the pseudobox labels, we proposed the weighted CIoU loss as is shown in (7), where O_{kij} denotes the NOM score, D_{kij} denotes the normalized distance score, and S_{kij} denotes the normalized size score of the jth proposed box of the ith point label of the kth image. By multiplying the NOM, normalized distance and normalized size score of all pseudolabels, respectively, to the IoU, distance, and aspect ratio term of CIoU loss, our weighted loss neutralizes the side effect introduced by pseudocenter point and side length. In addition, a hyperparameter λ is introduced to multiply the distance and aspect loss for adjusting the significance of these two losses as

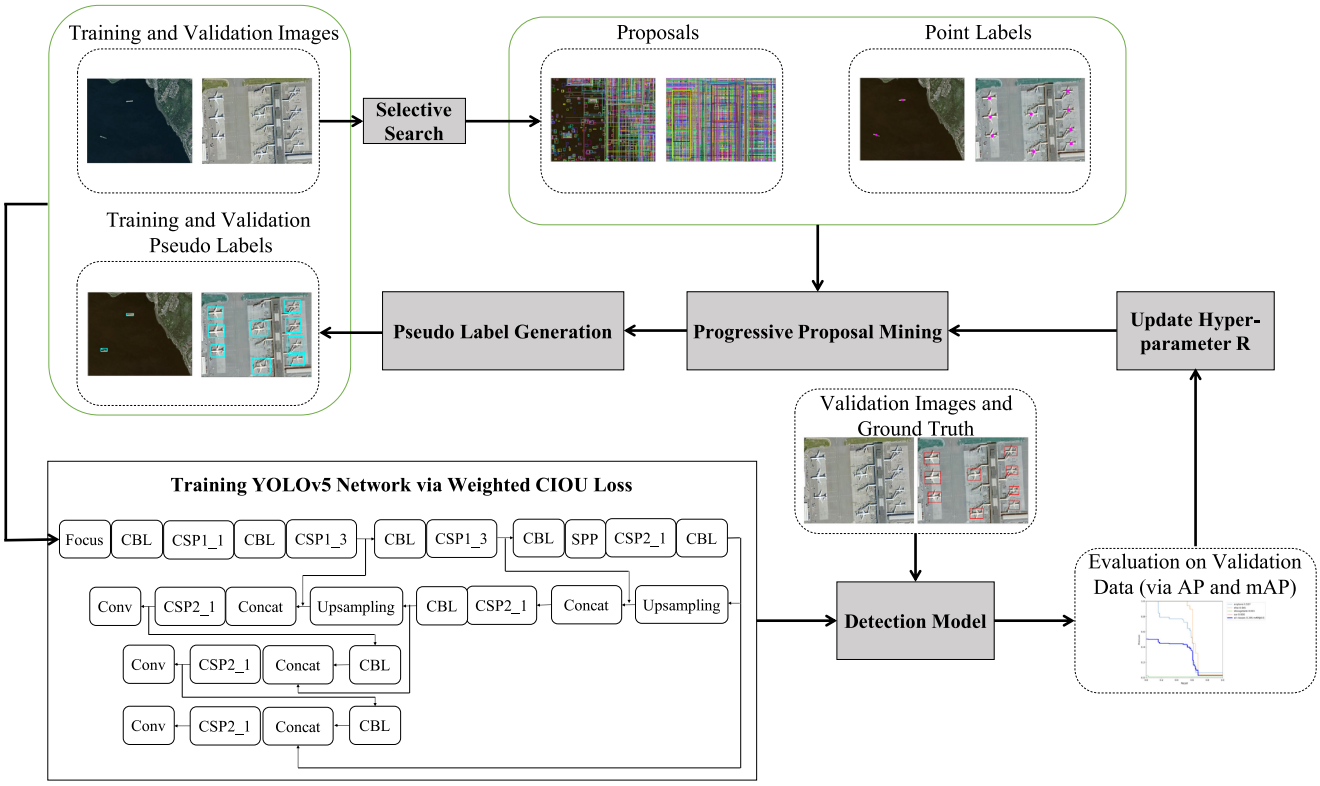

Fig. 3. Overview of the proposed method.

pseudobox labels are biased

$$WL_{\text{CIOU}} = 1 - \frac{IoU}{N \times n \times m} \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{m} O_{\text{kij}} + \frac{\rho^2(b, b^{gt})}{c^2} \times \frac{\lambda}{N \times n \times m} \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{m} D_{\text{kij}} + \frac{\lambda \times \alpha \times v}{N \times n \times m} \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{m} S_{\text{kij}}. \quad (7)$$

3) Progressive Proposal Mining: In order to optimize the value of the hyperparameter r, we propose a progressive refinement strategy for proposal mining. The first step is to randomly initialize the hyperparameter r between 0 and 1, and to initialize the hyperparameter $step$ which should be greater than 0 and less than the value of r. Then, the initial pseudoground truth label is generated from the top r proposals by (5). After training, the mean average precision (mAP) of predicted bounding box of validation data is evaluated, and update r by $step$. Repeat these two steps until the mAP of predicted bounding box of validation data stops increasing. The process of the progressive refinement strategy is shown in Algorithm 1. The overview structure of the proposed method is shown in Fig. 3.

III. Experiments

A. Dataset Description

We implemented our point-based WSL method to detect objects: Airplanes, ships, storage tanks, and cars in HSR remote sensing imagery. HSR remote sensing images of all classes are collected from a public dataset named NWPU VHR-10 dataset [27]–[29] that were cropped from Google Earth and Vaihingen dataset. The spatial resolution of images ranges from 0.08 to 2 m. Point annotations of images were manually annotated by experts. Samples of the dataset are shown in Fig. 4. There are overall 233 images. We divide them into a training dataset which consists of 163 images and corresponding point labels, a

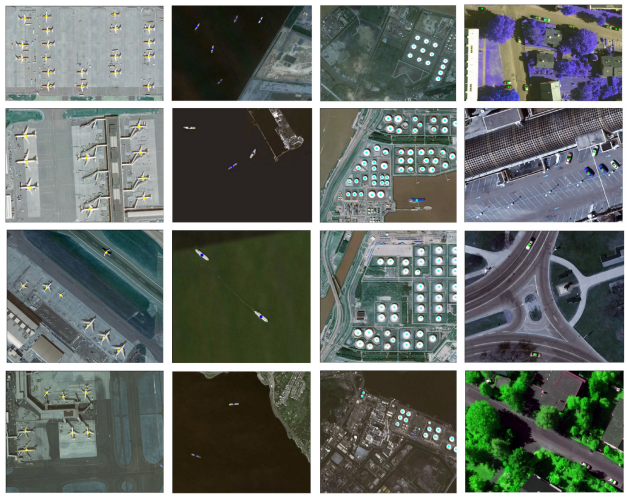

Fig. 4. Examples of images and point labels of the dataset. The first, second, third, and fourth columns represent airplane, ship, storage tank, and vehicle images and their corresponding point labels, respectively.

validation data which consists of 23 images and corresponding point and box labels, and a test data which consists of 47 images and corresponding box labels. It is noteworthy here that the box labels of validation and test data are only used for model evaluation.

B. Experimental Setup

We first generated around 10 000 proposals for each image from SS algorithm. Then, we conducted our proposal mining method on proposals, and generated initial pseudolabels. We utilized YOLOv5 as our backbone network, and pretrained the backbone network on the COCO dataset [30]. The convolution layers of the pretrained network were freezed, and the yolo layers were trained using our training data and related initial pseudolabels. In test and detection processes, NMS [31] is implemented to reduce duplicated bounding boxes with 0.2 IoU threshold. We evaluate our model by measuring the average precision (AP) and mAP. Note when the IoU between the ground truth and inference boxes is greater than 0.5 the inference boxes are treated as positive detection, which is consistent with the PASCAL VOC criteria [32].

C. Results

1) Normalized Overall Measure Analysis: Fig. 5 shows examples of pseudolabels generated from (5) of five categories. NOM represents the normalized overall measure value of pseudolabels, and r represents the ranking ratio of the NOM values of all candidate proposals, which means that the proposals of the highest r ratio are applied to generate pseudolabels. When r is smaller, the NOM score of pseudolabels is higher. Each column represents the pseudolabel of different categories corresponding to the same NOM and r. Each row represents the pseudolabel of the same categories generated from different NOM and r. Here, we show pseudolabels corresponding to r from 1 to 0.1 at intervals of 0.1. From Fig. 5, as the value of r decreases, the NOM increases, and the size of the pseudobounding box also decreases for all categories, which means the size of pseudolabels becomes smaller as the NOM score becomes higher. For the airplane category, when r is less than 0.3 NOM is greater than 0.77, the size of various pseudolabels is smaller than that of the airplane target, and the pseudolabels incline to fail to frame the wings of airplanes. This means that when the value of r is small the value of NOM is relatively big. The pseudolabel size tends to be smaller than the actual object size. We found that when NOM $= 0.77$ and $r = 0.3$, the size of the pseudolabels is appropriate for the airplane and ship categories in Fig. 5, but the label size is too large for the storage tank and vehicle categories.

2) Weighted Loss Analysis: Fig. 6 shows detailed detection performance on validation data implementing weighted loss with different λ. For all classes, when $\log_2^\lambda$ is greater than 3, the proposed weighted loss yields better performance than the baseline. This means when distance and aspect loss has a relatively bigger weight the model can provide a better performance. When $\log_2^\lambda$ equals to 5, the model obtained the best mAP@0.5. Thus, in the following experiments, the hyperparameter λ is set to 32. Regarding airplane class, λ has very limited impact on it, and its mAP@0.5 is near to 1, which means the generated pseudolabels of airplane are similar to the ground truth labels. Regarding ship class, when $\log_2^\lambda$ is less than 4, λ has no clear impact on ship detection performance, while when $\log_2^\lambda$ is greater than 4, λ has a negative impact on ship detection performance. In Fig. 5, the size and shape of ship pseudolabels are better than that of other classes especially storage tank and vehicle categories. Thus, if λ increases, the penalty of distance and side length also increases. This brings a negative impact for ship detection. For both storage tank and vehicle classes, bigger penalty on distance and side length can improve the performance of detection, since the pseudolabels of them are more diverse than those of the ship class.

To further illustrate the role of the proposed measurements, we studied the performance of the weighted loss without our measurements. In Fig. 7, detection models trained from different λ without scores were evaluated on validation data. Regarding all classes, the values of mAP@0.5 of all λ are almost equal to that of the baseline. Though λ is multiplied to loss, the λ constant does not change the essence of the loss. The performance of the airplane class is similar to Fig. 6. For the ship class, storage tank, and vehicle classes, there is a performance tradeoff between them. This results in the stable performance of all classes.

3) Proposal Mining Analysis: As described above, proposals are mined by (5). The proposal mining process and the impact of hyperparameter r were analyzed in this part. From Fig. 5, when r is lower than 0.5, the size of pseudolabels is much bigger than objects for all categories. Here, the value of r starts from 0.5, and setstep $= 0.1$ if $r \geq 0.1$, or step $= 0.02$ if $r < 0.1$. For each value of r, the corresponding pseudolabels were generated for training model. Then, the model was evaluated on validation data. The mining process stopped when the detection performance of validation data began to deteriorate. The hyperparameter r corresponding to the best detection performance of validation data was utilized to generate the final pseudo label.

Fig. 8 reports mAP@0.5 of all classes and shows precision recall curve (PRC) of models corresponding to r which is, respectively, equal to [0.5, 0.4, 0.3, 0.2, 0.1, 0.08, 0.06, 0.04]. When $r = 0.5$, the mAP@0.5 of airplane is 0.527 and the mAP@0.5 of ship is 0.641, while the mAP@0.5 for storage tank and vehicle classes is near to zero. The precision of storage tanks and cars are almost zero, which means the model cannot detect these two classes under $r = 0.5$. When $r = 0.4$, the mAPs@0.5 of airplane, ship, and storage tank are 0.772, 0.754, and 0.115, respectively, while the mAP@0.5 of vehicle class is near to zero. From the PRC, the curves of the car and storage tank classes are much lower than that of airplane and ship classes. When $r = 0.3$, the mAP@0.5 of storage tank class improved by 0.501, meanwhile the PRC curve of storage tank class rises a lot. When $r = 0.2$, the mAP@0.5 of car class improved by 0.158. The mAP@0.5 of storage tank class improved by 0.183. When $r = 0.1$, the mAPs@0.5 of storage tank and car class further improved, respectively, by 0.183 and 0.195. When $r = 0.08$, the mAPs@0.5 of all classes are slightly improved by 0.02. When $r = 0.06$, the mAPs@0.5 of car class significantly improved by 0.470. This improvement has a big impact on the mAP@0.5 of all classes, which also improved by 0.120. When $r = 0.04$, the

Fig. 5. Examples of pseudolabels for airplane, storage tank, vehicle, and ship generated from different hyperparameter r and NOM value.

mAPs@0.5 of ship, storage tank, and car classes have slightly dropped. By comparing all PRCs of different r values, the mAP@0.5 of all classes is gradually increasing until $r = 0.04$. The trend of car and ship classes is similar as other classes. The airplane class got its maximum mAP@0.5 when $r = 0.3$. When r further decreases, the mAP@0.5 of airplane class keeps stable. The storage tank class got its maximum mAP@0.5 when $r = 0.1$. When r further decreases, the mAP@0.5 of storage tank class keeps stable. From the analysis, we found the best r for separate class is different. However, when a class got its best performance, the performance can keep stable when r keep decreasing. When $r = 0.06$, the model obtained its best performance on all classes. Therefore, in the following experiments, the hyperparameter $r = 0.06$ was implemented to generate pseudolabels.

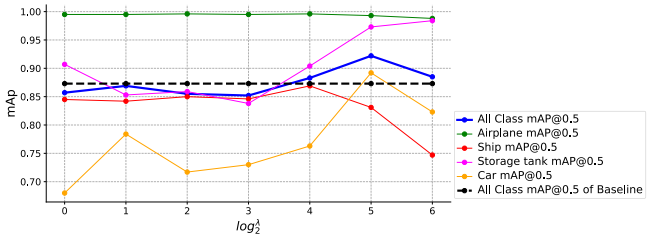

Fig. 6. Detection performance on validation data with weighted loss.

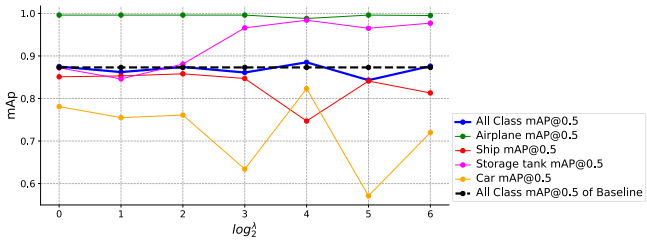

Fig. 7. Detection performance on validation data without weighted loss.

TABLE I
PERFORMANCE COMPARISONS (AP AND MAP) AMONG DIFFERENT OBJECT DETECTION METHODS

Method	Supervision	Airplane	Ship	Storage Tank	Car	mAP
YOLOv5	Box	0.996	0.955	0.979	0.928	0.965
WSDDN	Image-label	0.501	0.482	0.416	0.131	0.383
Ours	Point	0.969	0.914	0.972	0.839	0.924

4) Comparisons With Reference Methods: We compared our method with two state-of-the-art object detection algorithms. Yolov5 is a fully supervised method which is supervised by bounding boxes generated from experts, and WSDDN is a weakly supervised method which is supervised by image-level labels. Table I quantitatively evaluates the performance of YOLOv5, WSDDN and our method. From Table I, the fully supervised YOLOv5 obtained the highest AP and mAP among three methods. Our method achieved second high AP and mAP score. For the airplane category, our method is 0.027 lower than YOLOv5 in AP score, but 0.468 higher than WSDDN method. For the ship category, our method is 0.041 lower than YOLOv5 in AP score, but 0.432 higher than WSDDN method. For storage tank category, our method is 0.007 lower than YOLOv5 in AP score, but 0.556 higher than WSDDN method. For car category, our method is 0.089 lower than YOLOv5 in AP score, but 0.708 higher than WSDDN method. For all classes, our method is 0.041 lower than YOLOv5 in AP score, but 0.541 higher than WSDDN method. By implementing point supervision, our method has significantly shortened the time to make labels but achieved almost the same outcome as the fully supervised method. Compared to image-level supervision method, our method has achieved a substantial improvement in performance on the test set. Fig. 9 shows some examples of inference results on the test set. In general, our method accurately identified most of the targets, but there are still a small number of omissions and false positives. It is worth mentioning that in the third picture of the first row, the airplane with only the wings was also detected, indicating that our method is robust to the shape of the object. In the fifth picture on the first row, although the background of the image is very complicated, most of the airplanes are recognized. In the sixth picture of the first row, many targets are obscured by shadows and trees, and the locations of the cars are dense and irregular. However, our method also accurately identifies most cars. Most of the unidentified cars are obscured by shadows or trees. In the second line of the picture, almost all dense storage tanks are accurately identified one by one. In the second line of the figure, almost all dense storage tanks are accurately identified one by one. It is worth mentioning that in the second picture in the second row, some gray circular targets were mistakenly identified as storage tanks. However, it is difficult for the human eyes to determine whether these gray targets are storage tanks or not solely based on image information. In the third row of the figure, almost all ships have been successfully identified. In the fourth line of the image, we can find that no matter what color the car is, it can almost always be recognized. Especially in the last picture in the fourth row, despite the fact that the shadows obscure many cars, those cars are still successfully identified. Even a few cars that are completely obscured by shadows that are not easily recognized by the human eye are accurately identified by our method. This once again demonstrates that our algorithm is robust to shapes.

TABLE II
COMPLEXITY COMPARISON BETWEEN METHODS

	WSDDN	YOLOv5	Proposed
Average Annotation Time (seconds/per image)	3.18	54.51	14.22
Average Training Time (seconds/per epoch)	43.35	5.54	5.63
Average Inference Time (seconds/per image)	0.16	0.01	0.01

5) Complexity Analysis: In this part, annotation, training, and inference times of each method are compared to comprehensively analyze their efficiency. To evaluate annotation time, we randomly choose 100 images from our dataset. Then, we separately counted the overall time required to mark the image category label, bounding box label, and point label. Finally, we divided the total time by 100 to get the average manual labeling time required for each image. As shown in Table II, as WSDDN only needs to point categories of objects within each image, WSDDN demands the least time to label images; YOLOv5 needs the longest time to label bounding boxes for images; the annotation time of the proposed point-based WSL method is 11.04-s longer than WSDDN, but 40.29-s shorter than YOLOv5. This shows that compared with the reference supervision method, the proposed point-based WSL method greatly reduces the time of manual data labeling.

To evaluate the time complexity of these methods, we tested them on NVIDIA GeForce RTX 2070. Training time refers to 50 epochs of training for all methods. It is normalized to the total number of epochs to get the average training time per epoch. The inference time of all test images is recorded first, and then divided by the number of test images to get the average inference time for each image. As shown in Table II, compared to YOLOv5 and the proposed method, WSDDN requires the longest training and inference times; YOLOv5 and the proposed method require almost the same training and inference times. This means that the

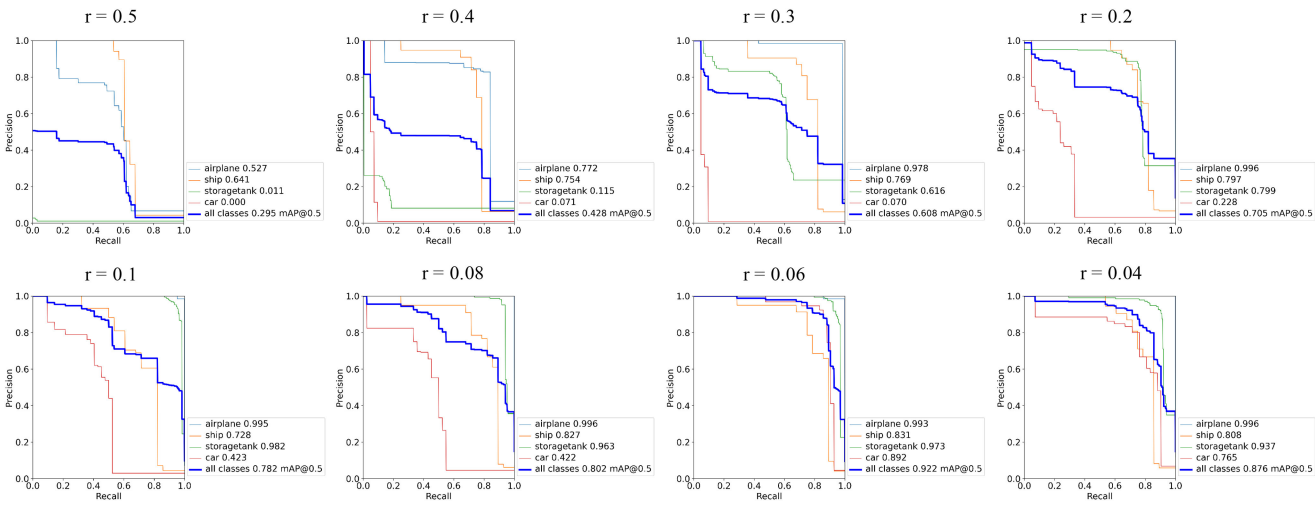

Fig. 8. Figure illustrates the precision recall curve of models trained by pseudolabels generated from different hyperparameter r on validation data.

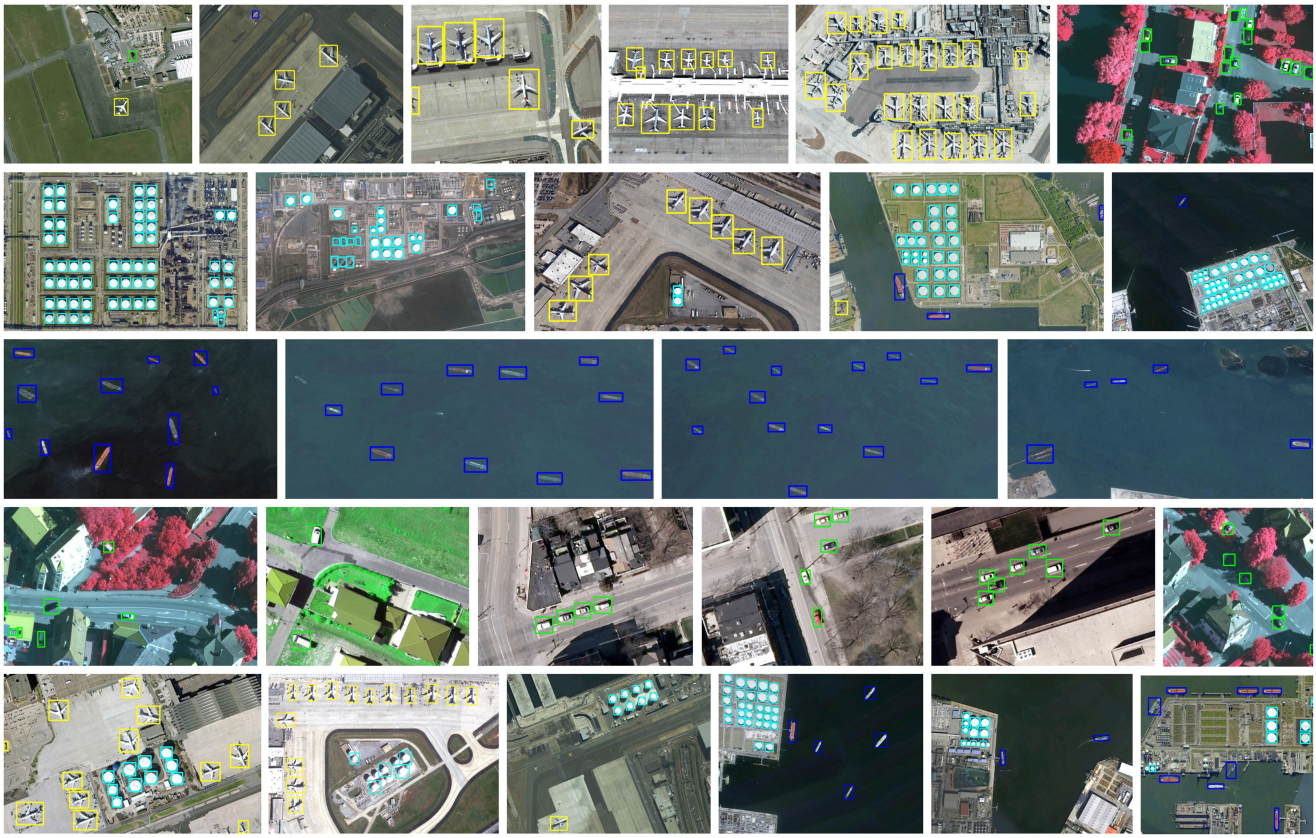

Fig. 9. Example results on the test split for each class. Yellow rectangle indicates airplane class; cyan rectangle indicates storage tank class; blue rectangle indicates ship class; and green rectangle indicates vehicle class.

proposed method and YOLOv5 are more efficient than WSDDN during training and inference.

IV. Conclusion

In this article, a novel point-based weakly supervised learning method is proposed to address object detection tasks in HSR remote sensing images. First, bounding box proposals are obtained from an unsupervised SS method. Three normalized measurements are introduced to evaluate the performance of proposals. Then, proposals are progressively mined to generate pseudobounding box labels depending on the performance of validation data. To train detection models, a weighted CIoU loss is therefore proposed balancing the uncertainty of pseudolabels.

We assess and analyze the proposed scheme on four classes which are airplane, ship, storage tank, and car. The results are compared to two state-of-the-art methods which are a fully supervised learning method and a WSL method. Our method achieves competitive performance compared to the fully supervised learning method, while our method greatly reduces human intervention. In addition, the performance of our method largely outperforms that of the image-level based on WSL method. The results point out that our method is a useful alternative for object detection in HSR remote sensing images.

In this research, our point-based weakly supervised method mainly focuses on identifying objects in HSR remote sensing images. In future research, we also hope to work on recognizing objects at other scales. Also, we hope to further improve the object recognition accuracy of weakly supervised learning based on point labels and reduce manual intervention and interference while maintaining state-of-the-art localization and detection accuracy. The proposed point-based WSL method is not only applicable for object detection in RGB images but also suitable for other types of remote sensing data such as multispectral, hyperspectral, and SAR image data. However, there might be some difficulties to be solved in the generalization to these types of data. Especially for SAR data, unsupervised clustering method can hardly generate meaningful proposals for objects, since SAR images are inherently affected by speckle noise, and the visual interpretability of SAR images is not as natural as in optical images. Since the proposal generation process is usually an important but time-consuming stage for WSL methods, we envision to focus on strategies for effectively generating proposals in our future research.

REFERENCES

[1] V. S. Martins, A. L. Kaleita, B. K. Gelder, H. L. da Silveira, and C. A. Abe, "Exploring multiscale object-based convolutional neural network (multi-ocnn) for remote sensing image classification at high spatial resolution," *ISPRS J. Photogrammetry Remote Sens.*, vol. 168, pp. 56–73, Oct. 2020.

[2] X. Qi *et al.*, "MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS J. Photogrammetry Remote Sens.*, vol. 169, pp. 337–350, Nov. 2020.

[3] G. D. Martins, M. d. L. B. T. Galo, and B. S. Vieira, "Detecting and mapping root-knot nematode infection in coffee crop using remote sensing measurements," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 12, pp. 5395–5403, Aug. 2017.

[4] I. W. Housman, R. A. Chastain, and M. V. Finco, "An evaluation of forest health insect and disease survey data and satellite-based remote sensing forest change detection methods: Case studies in the United States," *Remote Sens.*, vol. 10, no. 8, Aug. 2018, Art. no. 1184.

[5] N. Kabisch, P. Selsam, T. Kirsten, A. Lausch, and J. Bumberger, "A multi-sensor and multi-temporal remote sensing approach to detect land cover change dynamics in heterogeneous urban landscapes," *Ecological Indicators*, vol. 99, pp. 273–282, Apr. 2019.

[6] Q. Yao, X. Hu, and H. Lei, "Multiscale convolutional neural networks for geospatial object detection in VHR satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 23–27, Feb. 2020.

[7] R. Dong, D. Xu, J. Zhao, L. Jiao, and J. An, "Sig-NMS-based faster R-CNN combining transfer learning for small target detection in VHR optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8534–8545, Jul. 2019.

[8] Y. Bazi and F. Melgani, "Convolutional SVM networks for object detection in UAV imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3107–3118, Feb. 2018.

[9] Y. Li, F. Melgani, and B. He, "CSVM architectures for pixel-wise object detection in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6059–6070, Mar. 2020.

[10] A. Movia, A. Beinat, and F. Crosilla, "Shadow detection and removal in RGB VHR images for land use unsupervised classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 119, pp. 485–495, Sep. 2016.

[11] Y. Zhong, X. Han, and L. Zhang, "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 138, pp. 281–294, Apr. 2018.

[12] Y. Gong *et al.*, "Context-Aware convolutional neural network for object detection in VHR remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 34–44, Sep. 2019.

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

[14] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[15] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Adv. Neural Inf. Process. Syst.*, vol. 10, pp. 570–576, Jul. 1998.

[16] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," *Adv. Neural Inf. Process. Syst.*, vol. 15, pp. 577–584, Dec. 2002.

[17] 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, Jul. 2015.

[18] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans Pattern Anal Machine Intell.*, vol. 39, no. 1, pp. 189–203, Feb. 2016.

[19] 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.*, 2015, pp. 1224–1232.

[20] 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.*, 2017, pp. 2843–2851.

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

[22] 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, May 2020.

[23] P. Mettes and C. G. Snoek, "Pointly-supervised action localization," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 263–281, Mar. 2019.

[24] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.

[25] G. Jocher *et al.*, "Ultralytics/YOLOv5: V3.1 - Bug fixes and performance improvements," Oct. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4154370

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

[27] 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. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.

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

[29] 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, Sep. 2016.

[30] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput Vis.*, Sep. 2014, pp. 740–755.

[31] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4507–4515.

[32] 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. Machine Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

Youyou Li received the B.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2014, where she is currently working toward the Ph.D degree in remote sensing with the School of Resources and Environment and Center for Information Geoscience.

Her research interests include machine learning and pattern recognition on remote sensing data.

Binbin He (Member, IEEE) received the B.S. degree in resource exploration engineering and the M.S. degree in geology from the Chengdu University of Technology, Chengdu, China, in 1996 and 2002, and the Ph.D. degree in photogrammetry and remote sensing from China University of Mining and Technology, Xuzhou, China, in 2005.

He is currently a Professor with the School of Environment and Resources, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include quantitative estimation of land surface variables from satellite remote sensing, and spatiotemporal data mining for big data.

Farid Melgani (Fellow, IEEE) received the state engineer degree in electronics from the University of Batna, Batna, Algeria, in 1994, the M.Sc. degree in electrical engineering from the University of Baghdad, Baghdad, Iraq, in 1999, and the Ph.D. degree in electronic and computer engineering from the University of Genoa, Genoa, Italy, in 2003.

He is a Full Professor of Telecommunications with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, where he teaches pattern recognition, machine learning, and digital transmission. He is the Head of the Signal Processing and Recognition Laboratory, and the Coordinator of the Doctoral School in Industrial Innovation. He has coauthored more than 240 scientific publications. His research interests are in the areas of remote sensing, signal/image processing, pattern recognition, machine learning, and computer vision.

Dr. Melgani is currently an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *International Journal of Remote Sensing*, and IEEE JOURNAL ON MINIATURIZATION FOR AIR AND SPACE SYSTEMS.

Teng Long received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2012, 2015, and 2020 respectively.

His research interests include few-shot learning and non-Euclidean learning.