Learning Remote Sensing Object Detection With Single Point Supervision

Shitian He, Huanxin Zou, Yingqian Wang, *Member, IEEE*, Boyang Li, Xu Cao, and Ning Jing

***Abstract*— Pointly supervised object detection (PSOD) has attracted considerable interest due to its lower labeling cost when compared to box-level supervised object detection. However, the complex scenes and densely packed and dynamic-scale objects in remote-sensing (RS) images hinder the development of PSOD methods in the RS field. In this article, we make the first attempt to achieve RS object detection with single-point supervision and propose a PSOD method tailored for RS images. Specifically, we design a point label upgrader (PLUG) to generate pseudo-box labels from single-point labels and then use the pseudo-boxes to supervise the optimization of existing detectors. Moreover, to handle the challenge of the densely packed objects in RS images, we propose a sparse feature-guided semantic prediction (SemPred) module that can generate high-quality semantic maps by fully exploiting informative cues from sparse objects. Extensive ablation studies on the DOTA dataset have validated the effectiveness of our method. Our method can achieve significantly better performance when compared to state-of-the-art image-level and point-level supervised detection methods and reduce the performance gap between PSOD and box-level supervised object detection. The code is available at https://github.com/heshitian/PLUG.**

***Index Terms*— Remote sensing (RS), single-point supervised object detection (PSOD), sparse guided feature aggregation.**

I. Introduction

REMOTE-SENSING object detection (RSOD) plays an important role in many fields, such as national defense and security, resource management, and emergency rescuing. With the development of deep learning, many deep neural network (DNN)-based detection methods [1], [2], [3], [4], [5], [6], [7] were proposed and achieved promising performance. Besides, a number of remote-sensing (RS) datasets (e.g., HRSC2016 [8], NWPU VHR-10 [9], and DOTA series [10]) containing accurate and rich annotations were proposed to develop and benchmark RSOD methods. In these datasets, accurate location, scale, category, and quantity information of objects are provided and greatly facilitate the development of RSOD. However, such rich annotation formats will lead to expensive labor costs when RSOD methods are transferred to the new RS data (e.g., images captured by new satellites).

Manuscript received 24 November 2023; accepted 13 December 2023. Date of publication 18 December 2023; date of current version 4 January 2024. This work was supported by the National Natural Science Foundation of China under Grant 62071474. *(Corresponding author: Huanxin Zou.)*

The authors are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: heshitian19@nudt.edu.cn; zouhuanxin@nudt.edu.cn; wangyingqian16@nudt.edu.cn; liboyang20@nudt.edu.cn; cx2020@nudt.edu.cn; ningjing@nudt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3343806

To reduce the labor costs of annotating new RS data, researchers explored image-level annotations where only category information of objects is provided and introduced image-level supervised detection methods [11], [12], [13], [14], [15], [16]. These methods generally detect objects in a "find-and-refine" pipeline, that is, the coarse positions of objects are first found, and the proposals are then generated and refined. However, due to the complex RS scenes and the lack of location, scale, and quantity information, it is highly challenging to achieve good RSOD performance based on image-level annotation. Recently, single-point annotation [17], [18], [19], [20] has attracted much attention. Different from image-level annotations, point labels can simultaneously provide category, quantity, and coarse position information. The introduction of additional location and quantity information simplifies the original "find-and-refine" pipeline to the "refine-only" one and thus reduces the difficulties of pseudo-box generation. Besides, the labor cost of single-point annotations is only about 1/18 of box-level labels [19] and is negligibly higher than image-level ones. Therefore, single-point annotations have a large potential in the detection field.

Pointly supervised object detection (PSOD) is still in its infancy, with just a few methods [17], [18], [19] being proposed in recent years. Papadopoulos et al. [17] introduced center-click annotation and used the error distribution between two clicks to estimate object scales. Ren et al. [18] proposed a unified object detection framework that can handle different forms of supervision (e.g., tags, points, scribbles, and boxes) simultaneously. Chen et al. [19] predefined massive proposals in varied scales, aspect ratios, and shaking degrees for each point label and used multi-instance learning (MIL) to select and refine the most suitable proposals as the final results.

A straightforward way to achieve pointly supervised RSOD is to directly apply existing PSOD methods to RS images. These PSOD methods mainly follow the MIL pipeline, in which many proposals are preset for each point label, and then the optimal one is selected as the pseudo-box label. However, this framework is unsuitable for the RSOD task due to the low recall of proposal bags caused by the extremely huge variation of scales and aspect ratios of RS objects. In this article, we make the first attempt to achieve RSOD with single-point supervision and propose a point label upgrader (PLUG) to generate high-quality pseudo-box labels from single points. Specifically, the semantic response map is first learned under point-level supervision, and then pseudo-boxes can be generated in the shortest path paradigm. Due to the

discard of proposal generation, our PLUG is less susceptible to interference from varied scales and aspect ratios. Moreover, the dense and cluttered objects in RS images hamper the extraction of discriminative features and thus degrade the qualities of generated pseudo-boxes. Considering this issue, we propose a sparse feature-guided semantic prediction (SemPred) module to extract general representations of sparse objects and utilize them to improve the quality of the pseudo-boxes of dense objects. In this way, our PLUG can obtain more discriminative feature representations and improve the downstream detection performance.

By utilizing PLUG to transform single-point labels into box-level ones, we can develop a PLUG-Det method to achieve PSOD tailored for RS images. The training pipeline of our PLUG-Det consists of three stages (as shown in Fig. 1). First, our PLUG is trained under single-point supervision. Then, pseudo-boxes are generated by performing inference using the well-trained PLUG. Finally, existing fully supervised detectors (e.g., Faster-RCNN) are trained using the pseudo-boxes to achieve PSOD.

In summary, our main contributions are as follows.

1) We present the first study on single-point supervised RSOD and propose a simple, yet effective method called PLUG to generate pseudo-box annotations from single-point ones.
2) To handle the challenge of dense and clustered objects in RS images, we propose a sparse feature-guided SemPred approach to enhance the discriminative feature representation capability of our PLUG.
3) By using the generated pseudo-boxes to train existing detectors (Faster-RCNN [21] in this article), our method (i.e., PLUG-Det) achieves promising detection performance and outperforms many existing weakly supervised detectors.

The remainder of this article is organized as follows. In Section II, we briefly review the related works. Section III presents the details of the proposed method. Comprehensive experimental results are provided in Section IV, and Section V concludes this article.

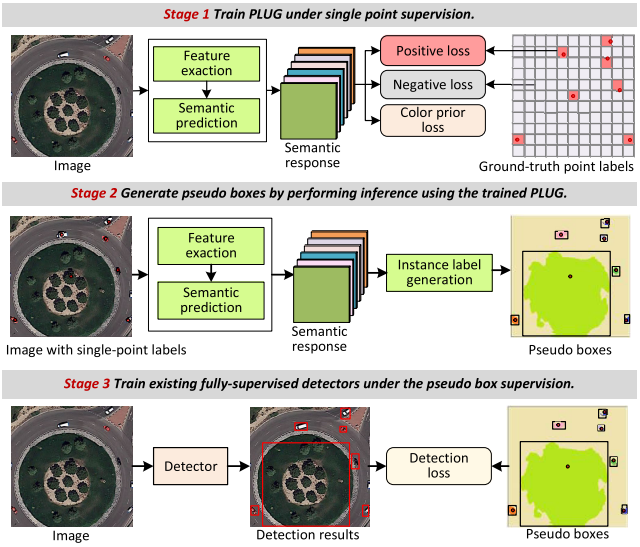


Fig. 1. Illustration of the training pipeline of our PSOD method. First, our PLUG is trained under single-point supervision. Then, pseudo-box labels are generated by performing inference using the well-trained PLUG. Finally, existing fully supervised detectors (e.g., Faster-RCNN) are trained under the supervision of the generated pseudo-boxes.

II. Related Works

A. Object Detection in Remote-Sensing Images

RSOD has been extensively investigated in the past decades. Since the convolutional neural network (CNN) was proposed, deep-learning-based RSOD methods have achieved significant improvements [22]. Compared to objects in natural images, RS objects have some special characteristics [23], including varied orientation, dynamic scales, densely packed arrangements, significant intraclass difference, and so on. Therefore, RSOD methods generally focus on the solutions to the above unique issues.

Specifically, regarding the varied orientation issue, many researchers proposed new representation approaches, for example, rotated bounding boxes [24], [25], [26], [27], intersecting lines [28], [29], key points [30], [31], [32], and rotated Gaussian distribution [33], [34]. Besides, some researchers proposed improved feature extraction modules [7], [35], [36], novel loss functions [23], [37], and new angle regression mechanisms [38], [39] to improve the detection performance on multioriented objects. Regarding the dynamic scales issue, Hu et al. [40] proposed a feature enhancement method that can extract more discriminative features containing structure, deep semantic, and relation information simultaneously. In [23] and [41], multiscale features were used to extract the scale-invariant representation of objects. Besides, Li et al. [6] proposed a ground sample distance (GSD) identification subnetwork and combined GSD information with the sizes of regions of interest (RoIs) to determine the physical size of objects. Regarding the densely packed arrangement issue, Yang et al. [42] proposed ClusDet, in which clustering regions were first extracted by a cluster proposal subnetwork, and then fed to a specific detection subnetwork for final prediction. Li et al. [43] proposed a density map-guided detection method, where the density map can represent whether a region contains objects or not and thus guide cropping images statistically.

Apart from the above studies, there are still many works trying to tackle other issues (e.g., excessive feature coupling [44], [45], unbalanced label assignment [46], and various aspect ratios [47], [48]) in RSOD. Recently, Transformer-based object detection methods [49], [50], [51] have attracted much attention due to their strong modeling capability. Therefore, some Transformer-based RSOD methods [52], [53] have been proposed and achieved remarkable detection performance.

The aforementioned methods improve the detection performance under box-level supervision. In this article, we aim to relieve the labor cost of annotating RS images and propose a single-point supervised RSOD method.

B. Image-Level Supervised Object Detection

To relieve the burden of box-level labeling, numerous image-level supervised detection methods [11], [12], [13],

[14], [15], [16], [54], [55], [56] were proposed, which can be categorized into class activation map (CAM)-based and MIL-based methods.

CAM-based methods [15], [16] detect objects based on the CAMs. Li et al. [57] proposed a CAM-based detection framework, in which the mutual information between images was exploited, and the class-specific activation weights were learned to better distinguish multiclass objects. Since CAM-based methods can only generate a few proposals for each class [58], it is not suitable for RS images with multiple instances.

MIL-based methods [11], [12], [13], [14] generally utilize off-the-shelf proposal generators (e.g., selective search [59], edge boxes [60], and sliding windows) to produce initial proposals and then consider the proposal refinement process as an MIL problem to make final predictions [58]. For example, WSDDN [11] first generates proposals using edge boxes and then feeds the extracted features of proposals to two parallel branches for classification and detection scoring, respectively. The two obtained scores are used to classify positive proposals. Based on WSDDN, OICR [12] uses selective search to generate proposals and adds an instance classification refinement process to enhance the discriminatory capability of the instance classifier. PCL [13] improves the original proposal bags to proposal clusters so that spatially adjacent proposals with the same label can be assigned to the same category cluster.

In 2014, Zhang et al. [61] first transferred the image-level supervised detection methods into the RSOD field. Specifically, they first performed saliency-based segmentation and negative sample mining to generate initial training samples and then proposed an iterative training approach to refine the samples and the detector gradually. On this basis, Han et al. [62] proposed a Bayesian framework to generate training samples, in which a deep Boltzmann machine was employed to extract the high-level features. In the image-level supervised RSOD field, the key challenging issues are local discrimination, multiinstances, and the imbalance between easy and difficult samples. Recent methods put efforts into the improvement of these issues. For example, regarding the local discrimination issue, Feng et al. [63] proposed a novel triple context-aware network, named TCANet, to learn complementary and discriminative visual features. Feng et al. [56] subsequently proposed a progressive contextual instance refinement method. Qian et al. [64] proposed a semantic segmentation-guided pseudo-label mining module to mine high-quality pseudo-ground-truth instances. Regarding the multi-instances issue, Wang et al. [65] proposed a unique multi-instance graph learning framework. Feng et al. [66] proposed to utilize the rotation consistency to pursue all possible instances. Wang et al. [67] developed a novel multiview noisy learning framework, named MOL, which uses reliable object discovery and progressive object mining to reduce background interference and tackle the multi-instance issue. For the imbalanced easy and difficult samples, Yao et al. [68] performed dynamic curriculum learning to progressively learn the object detectors in an easy-to-hard manner. Qian et al. [69] incorporated a difficulty evaluation score into training loss to alleviate the imbalance between easy and difficult samples.

The aforementioned studies improve the detection performance of image-level supervised RSOD methods. However, since image-level annotations cannot provide enough location and quantity information, these methods cannot achieve reasonable performance when applied to the RSOD task (see Section IV). In this article, we sacrifice little labor cost and focus on single-point supervised RSOD.

C. Point Supervision in Vision Tasks

Recently, point-level labels gradually attracted research attention due to their similar labeling time and richer labeling information. Point-level supervision has been extensively investigated in many vision tasks, including object detection [17], [18], [19], semantic segmentation [70], [71], [72], instance segmentation [73], [74], [75], panoptic segmentation [76], localization [20], [77], [78], infrared small target segmentation [79], [80], and so on.

Wu et al. [72] proposed a deep bilateral filtering network (DBFNet) for single-point supervised semantic segmentation, in which a bilateral filter was introduced to enhance the consistency of features in smooth regions and enlarge the distance of features on different sides of edges. Cheng et al. [74] proposed a multipoint supervised instance segmentation (PSIS) method, named Implicit PointRend, that can generate parameters of the mask prediction function for each object. Fan et al. [76] considered panoptic pseudo-mask generation as a shortest path searching puzzle and used semantic similarity, low-level texture cues, and high-level manifold knowledge as traversing costs between adjacent pixels. Yu et al. [20] proposed a coarse point refine (CPR) method for single-point supervised object localization, and the CPR method can select semantic-correlated points around point labels and find semantic center points through MIL learning.

In the object detection field, Papadopoulos et al. [17] first introduced center-click annotation, in which the error distribution between two clicks is utilized to estimate object scales. Hence, two repetitive and independent center annotations are needed in their method. Different from that, our method tries to generate pseudo-boxes from a single arbitrary point on the object mask. Ren et al. [18] proposed a unified object detection framework (i.e., UFO^2) that can handle different forms of supervision (e.g., tags, points, scribbles, and boxes) simultaneously. Different from handling different forms of supervision, the emphasis of our method is better at generating pseudo-boxes from single points based on the characteristics of RS objects. Chen et al. [19] proposed an MIL-based single-point supervised detection framework that can adaptively generate and refine proposals via multistage cascaded networks. In their method, proposal bags are generated through some fixed parameters that control the proposal scales, aspect ratios, shaking degrees, and quantities. However, due to the challenges in the RS field (as mentioned in Section I), their method suffers performance degradation when applied to RS images. In this article, we focus on the special challenges of RSOD and explore single-point supervised detection methods tailored for RS images.

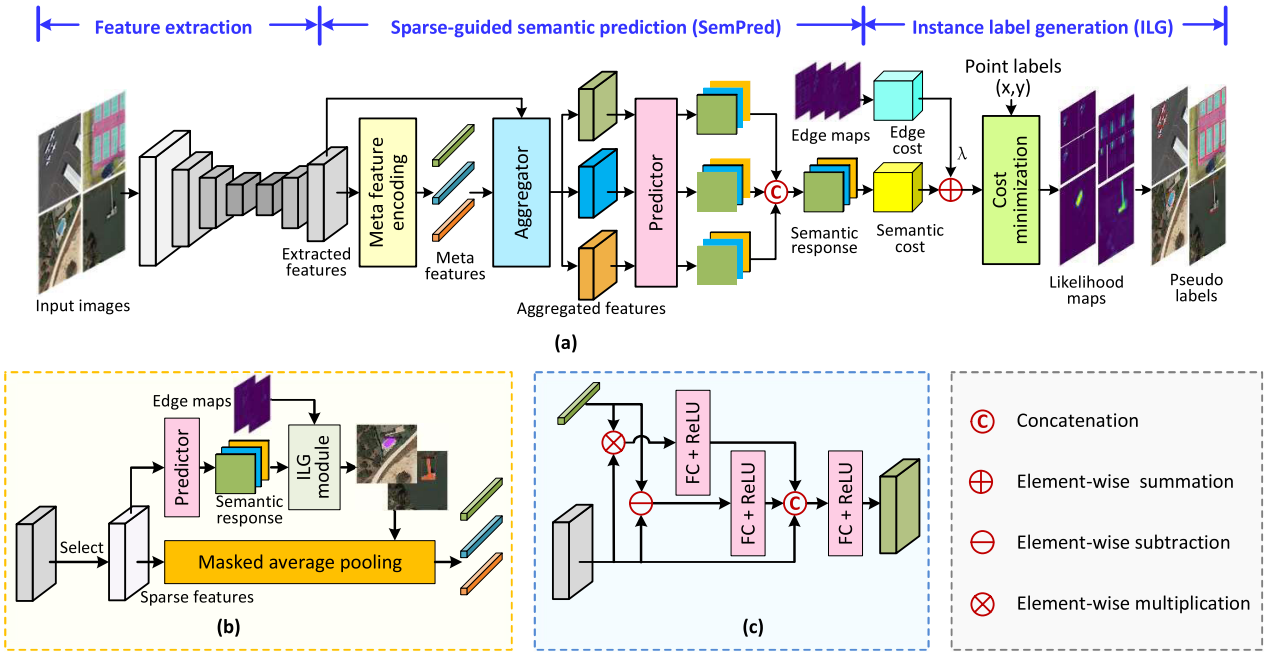


Fig. 2. Overview of the proposed PLUG, which is designed to transform point labels into pseudo-boxes. Specifically, the feature extraction module extracts discriminative features from input images. Then, the sparse feature-guided SemPred module takes the extracted features as its input and is responsible for the semantic response prediction. Finally, the ILG module takes both the input images and the predicted response as its input to generate pseudo-boxes. (a) Overall architecture. (b) Meta-feature encoding. (c) Aggregator.

III. Method

In this section, we introduce the details of our method. We first introduce the architecture of the proposed PLUG, which consists of the feature extraction module, the sparse feature-guided SemPred module, and the instance label generation (ILG) module (see Fig. 2). Afterward, we introduce the training losses of our PLUG.

A. Feature Extraction

In our method, ResNet [81] with feature pyramid network (FPN) [82] is used as the feature extraction module. The ResNet backbone extracts features of images of different scales, and FPN fuses the multiscale features to balance the contents of semantic and structural information. Following [20], the P2 layer (with 8× downsampling ratio) of FPN is used for subsequent processing.

B. Sparse Feature-Guided Semantic Prediction

Taking the extracted features as input, the sparse feature-guided SemPred module is responsible for obtaining the semantic response of objects, in which object regions are activated in the specific category layers. Besides, the SemPred module can reduce the difficulty of discriminative feature extraction on dense objects. Specifically, we observe that the pseudo-boxes generated on sparse objects are of higher quality than those generated on dense objects (see Section IV-F for details). Consequently, in our SemPred module, the general representation of sparse objects is used to enhance the extracted features and thus improve the discriminative feature representation capability of our PLUG. The detailed architecture of the SemPred module is shown in Fig. 2(a), which consists of three stages: meta-feature encoding, feature aggregation, and SemPred.

1) Meta-Feature Encoding: In this stage, the general representation (i.e., meta-feature) of sparse objects is encoded from the extracted features. As shown in Fig. 2(b), meta-feature encoding takes the extracted features as input and obtains sparse features by selecting the features of images with a single object. Then, the sparse features are fed to a predictor and the ILG module to generate the pseudo-labels of sparse objects. With the sparse features and the pseudo-labels, masked average pooling is performed to obtain the feature representation of each sparse object. To obtain more representative and stable meta-features, all the sparse representations in the dataset are averaged according to their categories. Finally, C (the number of categories) meta-features are obtained, each of which can represent the general information of objects in a specific category.

2) Feature Aggregation: After obtaining C meta-features, C aggregated features are generated in this stage by using meta-features to enhance the extracted features. The architecture of our aggregator is shown in Fig. 2(c). Specifically, for each meta-feature, element-wise subtraction and multiplication are first performed. Then, the processed features are concatenated with the original feature to obtain the aggregated features. Note that, a fully connected layer and a ReLU layer are used after each operation (i.e., subtraction, multiplication, and concatenation).

3) Semantic Prediction: For each aggregated feature, a predictor (composed of a linear layer and a Sigmoid function) is used for semantic response prediction. Since the representations in meta-features are category-aware, different aggregated

features are experts in predicting objects in corresponding categories. Hence, the specific layer of the semantic response from different aggregated features is selected and concatenated to generate the final semantic response. It is worth noting that the predictors in different branches and the meta-feature encoding module share the same architecture and parameters.

Note that, in the SemPred module, meta-feature encoding is performed in the training phase only. During inference, the meta-features have been optimized and stored in advance, and thus the extracted features can be directly aggregated. In fact, the guidance of sparse objects can be considered as a self-distillation process [83], where the sparse features are the *teacher* and can transfer knowledge (high-quality features) to the *student*. With the guidance of sparse objects, the semantic response can be enhanced and benefits the pseudo-box generation in the following ILG module.

C. Instance Label Generation

After obtaining the semantic response, the ILG module is designed to generate pseudo-box annotations. The core of this module is to assign each pixel to its most likely object or background. Based on the assignment results, we can obtain the bounding box of each object by finding the circumscribed rectangle of the corresponding pixels.

Specifically, let $\mathcal{L} = \{l_0, l_1, l_2, \ldots, l_L\}$ denote the set of instances, where l_0 denotes background and $\{l_1, l_2, \ldots, l_L\}$ denote L objects. Each pixel p on the image will be assigned to an instance according to

$$\mathcal{I}ns(p) = \arg\min_{l \in \mathcal{L}} \{\text{Cost}(p, p_l)\} \tag{1}$$

where p_l represents the point label of instance l that contains both location and instance information. $\text{Cost}(p, p_l)$ denotes the cost between pixel p and point label p_l. The core of the label assignment process in (1) is to find an instance with a minimum cost for each pixel.

The cost calculation between pixel p and point label p_l is formulated as a shortest path problem. Specifically, we formulate the cost between p and p_l as the second curvilinear curve integral along a given path $\Gamma \in \{\Gamma_1, \ldots, \Gamma_n\}$. That is,

$$\text{Cost}(p, p_l) = \min_{\Gamma \in \{\Gamma_1, \ldots, \Gamma_n\}} \int_{\Gamma} \left(C^{\text{sem}}(\vec{z}) + \lambda C^{\text{edge}}(\vec{z})\right) d\vec{z} \tag{2}$$

where $C^{\text{sem}}(\cdot)$ and $C^{\text{edge}}(\cdot)$ represent the semantic-aware neighbor cost and edge-aware neighbor cost, respectively, and λ is a hyperparameter to balance these two terms [76]. Specifically, $C^{\text{sem}}(\cdot)$ is the L_2 distance of the semantic response between two adjacent pixels. $C^{\text{edge}}(\cdot)$ is the L_1 distance of the edge map (generated by Sobel operator [84]) between two adjacent pixels, which can help better distinguish the densely packed objects (see Section IV-C3). Note that, $\text{Cost}(p, p_0)$ is manually set to a fixed threshold τ ($\tau = 0.5$ in our method) to assign pixels that are "far from" all the instances to the background. Besides, since there is no analytical solution to the integral in (2), we use Dijkstra's algorithm to obtain its numerical solution.

D. Losses

In the proposed PLUG, the ILG module is parameter-free, and the training process is only performed on the SemPred module. The losses to train the SemPred module have three parts including positive loss, negative loss, and color prior loss.

1) Positive Loss: Since point labels can provide accurate supervision on the annotated locations, we set these labeled pixels as positive samples and design a positive loss to optimize the SemPred module to generate correct predictions on these positions. The positive loss is designed based on the standard focal loss [85]

$$\mathcal{L}_{\text{pos}} = -\frac{1}{N_{\text{pos}}} \sum_{j=1}^{N_{\text{pos}}} \sum_{i=1}^{C} \Big[y_{ji} \left(1 - y'_{ji}\right)^{\gamma} \log\left(y'_{ji}\right) + \left(1 - y_{ji}\right) y_{ji}^{'\gamma} \log\left(1 - y'_{ji}\right) \Big] \tag{3}$$

where N_{pos} and C denote the total number of positive samples and categories, respectively. y and y' represent the ground-truth category label and the prediction scores, respectively. We follow the general settings in [85] to set γ to 2.

2) Negative Loss: In PSOD, only objects are labeled by single points, while the background regions are not annotated. Consequently, single-point annotations cannot provide sufficient supervision of the background. In our method, we follow this basic setting in PSOD and propose an approach to provide supervision on the background regions. Specifically, we suppose that background pixels are dominant in amount in the unlabeled region and then coarsely set all the unlabeled pixels as negative samples. Based on the coarse negative samples, we design a negative loss to enforce our model to better distinguish objects and background, that is,

$$\mathcal{L}_{\text{neg}} = -\frac{1}{N_{\text{neg}}} \sum_{j=1}^{N_{\text{neg}}} \sum_{i=1}^{C} \left(1 - y_{ji}\right) y_{ji}^{'\gamma} \log\left(1 - y'_{ji}\right) \tag{4}$$

where N_{neg} is the number of negative samples.

3) Color Prior Loss: We follow [76] to introduce a color prior loss, which can encourage adjacent pixels with similar colors to be classified to the same category, and enhance the prediction stability of our SemPred module. The color prior loss is formulated as

$$\mathcal{L}_{\text{col}} = -\frac{1}{Z} \sum_{i=1}^{HW} \sum_{j \in \mathcal{N}(i)} A_{i,j} \log y_i^{'T} y'_j \tag{5}$$

where y'_i and y'_j denote the category prediction scores of the ith and jth pixels, respectively. $A_{i,j}$ is the color prior affinity and is obtained by thresholding the pixel similarity computed in the LAB color space (with a threshold of 0.3). $\mathcal{N}(i)$ is the set of neighbor pixel indices of i. $Z = \sum_{i=1}^{HW} \sum_{j \in \mathcal{N}(i)} A_{i,j}$ is the normalization factor.

In summary, the overall loss is the weighted summation of the above three losses, that is,

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{pos}} + \alpha_1 \mathcal{L}_{\text{neg}} + \alpha_2 \mathcal{L}_{\text{col}} \tag{6}$$

where α_1, α_2 are two hyperparameters to balance different terms. In this article, α_1 and α_2 are set to $N_{\text{neg}}/N_{\text{pos}}$ and 1, respectively. With the well-designed loss function, our

PLUG can be well-optimized and effectively generate pseudo-bounding boxes.

IV. Experiments

In this section, we first introduce the datasets and implementation details and then combine the proposed PLUG with Faster-RCNN [21] to develop a PSOD method (i.e., PLUG-Det). Afterward, we compare PLUG-Det with image-level, point-level, and box-level supervised object detection methods. Moreover, we conduct ablation studies and make deep analyses to validate the effectiveness of our method. Finally, we develop a PLUG-Seg network by combining PLUG with Mask-RCNN [87] and conduct experiments to show the potential of our method in a single PSIS.

A. Datasets and Implementation Details

1) Datasets: To verify the effectiveness of our method, we conduct extensive experiments on the DOTA-v1.0 dataset [10], which contains 2806 large-scale RS images with 15 object categories, including plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HB), swimming pool (SP), and helicopter (HC). Objects in the DOTA dataset are labeled with box annotations. Since the iSAID dataset [88] contains the corresponding mask labels of objects in the DOTA dataset, we randomly selected a point on the mask of each object as the ground-truth point label. We used the training set and validation set for model development and performance evaluation, respectively. Due to hardware memory limitation, we cropped the original images into 512 $\times$ 512 patches with 128 overlapped pixels, and used the cropped patches for training and inference. In the training phase, random flip was used for data augmentation.

2) Implementation Details: We implemented our method based on the MMDetection [89] toolbox with an NVIDIA RTX 3090Ti GPU. The training of our PLUG-Det method consists of three stages: the training of PLUG, the inference of PLUG, and the training of existing detectors (e.g., Faster-RCNN). In the first stage, the learning rate was initially set to 0.001 and decreased by a factor of 0.1 at the 8th and 11th epoch, respectively. We trained our PLUG for a total of 12 epochs with a batch size of 8. Besides, we used the stochastic gradient descent (SGD) algorithm [90] for optimization. In the second stage, pseudo-boxes of the training set were obtained by performing inference using the trained PLUG. In this stage, the batch size was set to 1. In the third stage, we adopted the existing detector by default without modifying its hyperparameters. Taking Faster-RCNN with ResNet50 as an example, the learning rate was initially set to 0.005, and the optimizer was SGD with a 1$\times$ training schedule. Other training settings were kept as the default values in MMDetection [89]. The training times of the three stages are 4.8, 6, and 3.1 h, respectively. The total training time is the summation of the time spent in each stage and is about 14 h.

3) Evaluation Metrics: We used mean Intersection over Union (mIoU) between generated pseudo-boxes and ground-truth boxes to evaluate the performance of PLUG. Besides, mIoU_s, mIoU_m, and mIoU_l were used as the indicators to evaluate the quality of pseudo-boxes on small, medium, and large objects, respectively. Moreover, we evaluated the performance of PLUG-Det and its variants by reporting the mAP_{50} (averaged over IoU values with the threshold being set to 0.5) for all categories and the AP_{50} for each category. Similarly, mAP_s, mAP_m, and mAP_l were used to evaluate the detection performance on small, medium, and large objects, respectively.

B. Comparison to the State-of-the-Art Methods

In this section, we use the pseudo-boxes generated by different methods to train a Faster-RCNN detector and compare the detection performance of our PLUG-Det with existing image-level supervised and single-point supervised detection methods. Moreover, Faster-RCNN with ground-truth box-level supervision is also included to provide upper-bound results for reference.

Table I shows the AP_{50} values achieved by different detection methods. It can be observed that image-level supervised detectors (i.e., WSDDN [11], OICR [12], and OICR-FR [12]) achieve very low detection accuracy. Compared to those detectors, PSOD methods achieve better detection performance due to the extra coarse position and quantity information introduced by point annotations. Specifically, P2BNet-FR achieves an mAP_{50} score of 0.156 and can further achieve a 0.029 improvement with a two-stage cascaded optimization pipeline. In contrast, our PLUG-FR achieves an mAP_{50} score of 0.423, which significantly outperforms P2BNet-FR. The experimental results demonstrate the superiority of our method when compared to the MIL-based methods. It is worth noting that Faster-RCNN developed under ground-truth box supervision can achieve an mAP_{50} score of 0.648. That is, our PLUG-FR can achieve 65.3% of the performance of box-level supervised Faster-RCNN [21], but with an 18$\times$ reduction in annotation cost.

Besides, our method can generalized to different downstream detectors. We additionally use the one-stage detector FCOS [86] and the transformer-based detector Deformable DETR [50] to validate the generalization capability of our method. As shown in Table I, PLUG-FCOS and PLUG-Deformable DETR can achieve 0.360 and 0.322 in terms of mAP_{50} and are 66.2% and 55.8% of the performance of each fully supervised detectors, respectively. The consistent performance ratios compared to respective fully supervised detectors demonstrate the generality of our method.

Fig. 3 shows the qualitative results on eight typical scenes achieved by different detection methods. It can be observed that our PLUG-Det can achieve better detection performance than other state-of-the-art image-level supervised and single-point supervised detectors, especially on challenging scenes. Specifically, image-level supervised detectors (e.g., OICR-FR) may bring false alarms (e.g., Scene C) and miss detection (e.g., Scene F) due to their insufficient supervision.

TABLE I

AVERAGE PRECISION SCORES ACHIEVED BY DIFFERENT DETECTION METHODS ON THE DOTA DATASET. HERE, DEF-DETR REPRESENTS DEFORMABLE DETR

Method	Supervision	Backbone	Categories															mAP_{50}
			PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HB	SP	HC	
WSDDN [11]	*Image*	VGG16	0.003	0.009	0.000	0.005	0.000	0.001	0.000	0.003	0.000	0.000	0.013	0.000	0.010	0.010	0.000	0.004
WSDDN [11]	*Image*	ResNet50	0.014	0.064	0.001	0.013	0.021	0.030	0.016	0.034	0.004	0.025	0.019	0.053	0.011	0.044	0.004	0.023
OICR [12]	*Image*	VGG16	0.007	0.100	0.000	0.116	0.037	0.101	0.023	0.089	0.000	0.056	0.145	0.000	0.042	0.036	0.000	0.050
OICR [12]	*Image*	ResNet50	0.047	0.104	0.007	0.042	0.022	0.061	0.022	0.068	0.031	0.044	0.096	0.102	0.061	0.047	0.016	0.051
OICR-FR [12]	*Image*	ResNet50	0.042	0.038	0.000	0.002	0.075	0.301	0.037	0.077	0.011	0.132	0.033	0.159	0.050	0.120	0.001	0.072
FCOS [86]	*Box*	ResNet50	0.800	0.504	0.296	0.212	0.603	0.796	0.821	0.914	0.452	0.612	0.407	0.460	0.751	0.213	0.313	0.544
def-DETR [50]	*Box*	ResNet50	0.799	0.576	0.377	0.491	0.600	0.772	0.843	0.924	0.414	0.624	0.457	0.396	0.721	0.455	0.324	0.577
Faster-RCNN [21]	*Box*	ResNet50	0.850	0.665	0.435	0.587	0.588	0.831	0.833	0.933	0.493	0.634	0.590	0.589	0.791	0.534	0.373	0.648
P2BNet-FR [19]	*Point*	ResNet50	0.061	0.063	0.111	0.260	0.266	0.066	0.368	0.016	0.051	0.270	0.049	0.272	0.105	0.386	0.001	0.156
P2BNet-FR* [19]	*Point*	ResNet50	0.016	0.002	0.118	0.168	**0.397**	0.073	0.246	0.017	0.190	**0.465**	0.009	0.518	0.060	0.358	**0.140**	0.185
PLUG-FCOS(ours)	*Point*	ResNet50	0.353	0.340	0.226	0.111	0.296	**0.685**	**0.603**	**0.874**	**0.246**	0.455	0.192	0.468	0.349	0.171	0.039	0.360
PLUG-def-DETR (ours)	*Point*	ResNet50	0.250	0.398	0.241	0.166	0.288	0.614	0.547	0.795	0.090	0.383	0.160	0.345	0.227	0.353	0.047	0.322
PLUG-FR (ours)	*Point*	ResNet50	**0.509**	**0.543**	**0.291**	**0.284**	0.248	0.672	0.436	**0.874**	0.214	0.462	**0.360**	**0.543**	**0.438**	**0.446**	0.086	**0.427**

* means that P2BNet is optimized in a two-stage cascaded manner.

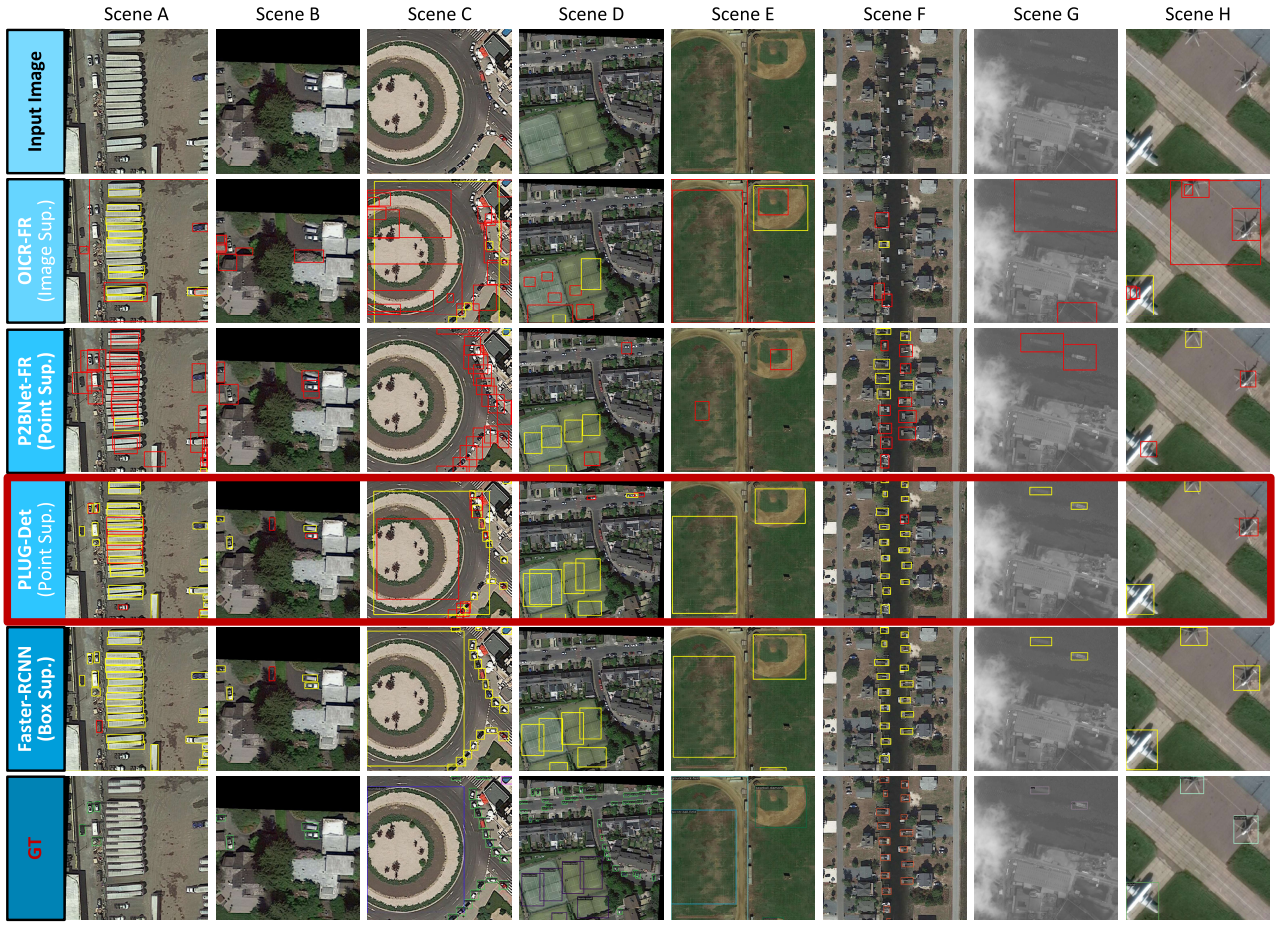


Fig. 3. Qualitative results obtained by different object detection methods on the DOTA validation set. The correctly detected results are marked by yellow boxes, and the falsely detected results are marked by red boxes. Gradually darker colors represent stronger supervision.

Besides, the single-point supervised detector P2BNet-FR has worse scale and aspect ratio adaptability compared with our method. For example, the vehicles in Scene A with large aspect ratios cannot be correctly detected by P2BNet-FR but can be better detected by our method.

C. Ablation Study

In this section, we conduct ablation studies to validate the effectiveness of our method.

1) Investigation of the Feature Extraction Module: We use ResNet [81] with FPN [82] as the feature extraction

TABLE II

COMPARISON OF THE PSEUDO-BOX QUALITY AND DETECTION PERFORMANCE ACHIEVED BY DIFFERENT BACKBONES. HERE, #PARAM REPRESENTS THE NUMBER OF PARAMETERS, AND FLOPS IS CALCULATED WITH A 512 × 512 INPUT IMAGE

backbone	FLOPs	#Param	Pseudo box quality				Detection performance			
			mIoU	$mIoU_s$	$mIoU_m$	$mIoU_l$	mAP_{50}	mAP_s	mAP_m	mAP_l
ResNet18	12.17 G	13.37 M	0.531	0.524	0.551	0.508	0.412	0.330	0.457	0.262
ResNet50	24.88 G	26.32 M	0.549	0.539	0.576	0.533	0.427	0.329	0.474	0.338
ResNet101	44.36 G	45.31 M	**0.558**	**0.548**	**0.584**	**0.564**	**0.436**	**0.338**	**0.490**	**0.384**

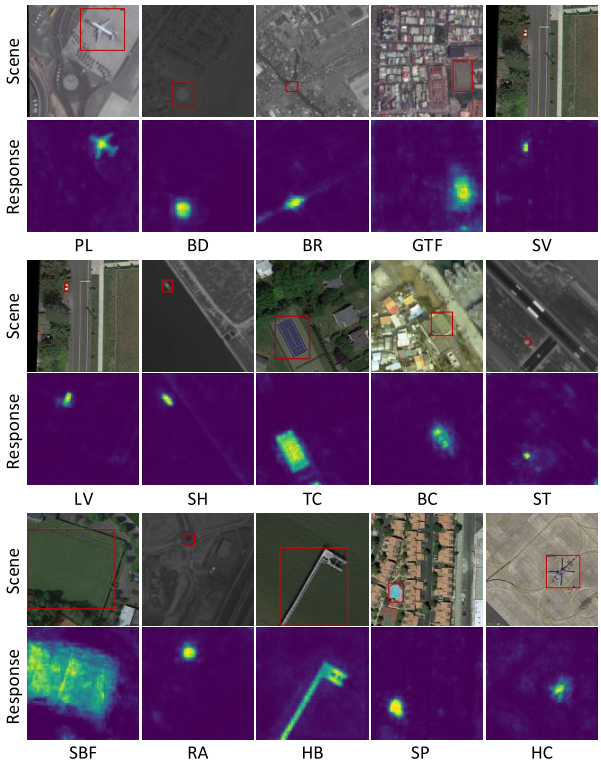


Fig. 4. Semantic response of images predicted by the SemPred module. Here, the layer of the corresponding category is visualized.

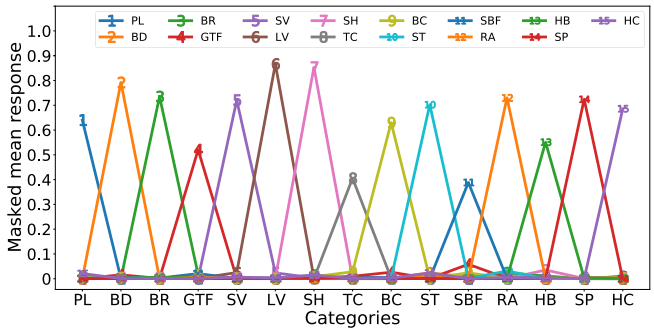


Fig. 5. Distribution of masked mean response in different categories.

module of our PLUG. Here, we compare the performance of our feature extraction module with different backbones (i.e., ResNet18, ResNet50, and ResNet101). We first evaluate the quality of generated pseudo-boxes on the training set. As shown in Table II, our PLUG can achieve mIoU scores of 0.531, 0.549, and 0.558 with ResNet18, ResNet50, and ResNet101 backbones, respectively. We also evaluate the downstream detection performance on the validation set. As shown in Table II, our PLUG-Det achieves an mAP_{50} score of 0.436 with ResNet101, which is higher than the mAP_{50} scores with ResNet18 and ResNet50. That is because the ResNet101 backbone is deeper and can extract more discriminative features. However, compared to ResNet18 and ResNet50, using ResNet101 as backbone introduces larger model size (1.82× of ResNet50) and higher FLOPs (1.78× of ResNet50). Consequently, we use ResNet50 as the backbone to achieve a good balance between accuracy and efficiency.

2) Effectiveness of the SemPred Module: The SemPred module utilizes meta-features of sparse objects to aggregate the extracted features and use the aggregated features for semantic response prediction. We conduct experiments to validate the effectiveness of the SemPred module and its key components.

Semantic Response Visualization: The semantic response prediction contains two potential tasks, including the recognition of objects from the background and the discrimination among categories. In this part, we validate the effectiveness of the SemPred module on these two tasks, respectively. First, we validate the object recognition capability by visualizing the predicted semantic response maps. As shown in Fig. 4, objects of different categories can be well distinguished from the background, and the response regions basically fit object shapes. Second, we validate the category discrimination capability of the SemPred module by visualizing the variation of masked mean response[1] on different category layers. As shown in Fig. 5, each object is only strongly activated on a single category layer. These results clearly demonstrate the effectiveness of the SemPred module in recognizing and classifying objects from backgrounds.

Sparse Feature Guidance (SFG): In the SemPred module, the general representations of sparse objects are used to aggregate the extracted features from backbones. To validate the SFG scheme, we replaced the SemPred module with a vanilla predictor (a linear layer followed by a Sigmoid function) and developed a variant (i.e., "vanilla" in Table III) of PLUG without the guidance of sparse objects. As shown in Table III, the mIoU score is improved from 0.497 to 0.549 when SFG is performed, and the mAP_{50} value of our PLUG-Det is also improved from 0.356 to 0.423 correspondingly. It demonstrates that the proposed SFG scheme can improve the quality of generated pseudo-boxes and thus benefits the downstream detection performance. Moreover, we compare the semantic

[1]Masked mean response denotes the average value of response map on the ground-truth mask of each object.

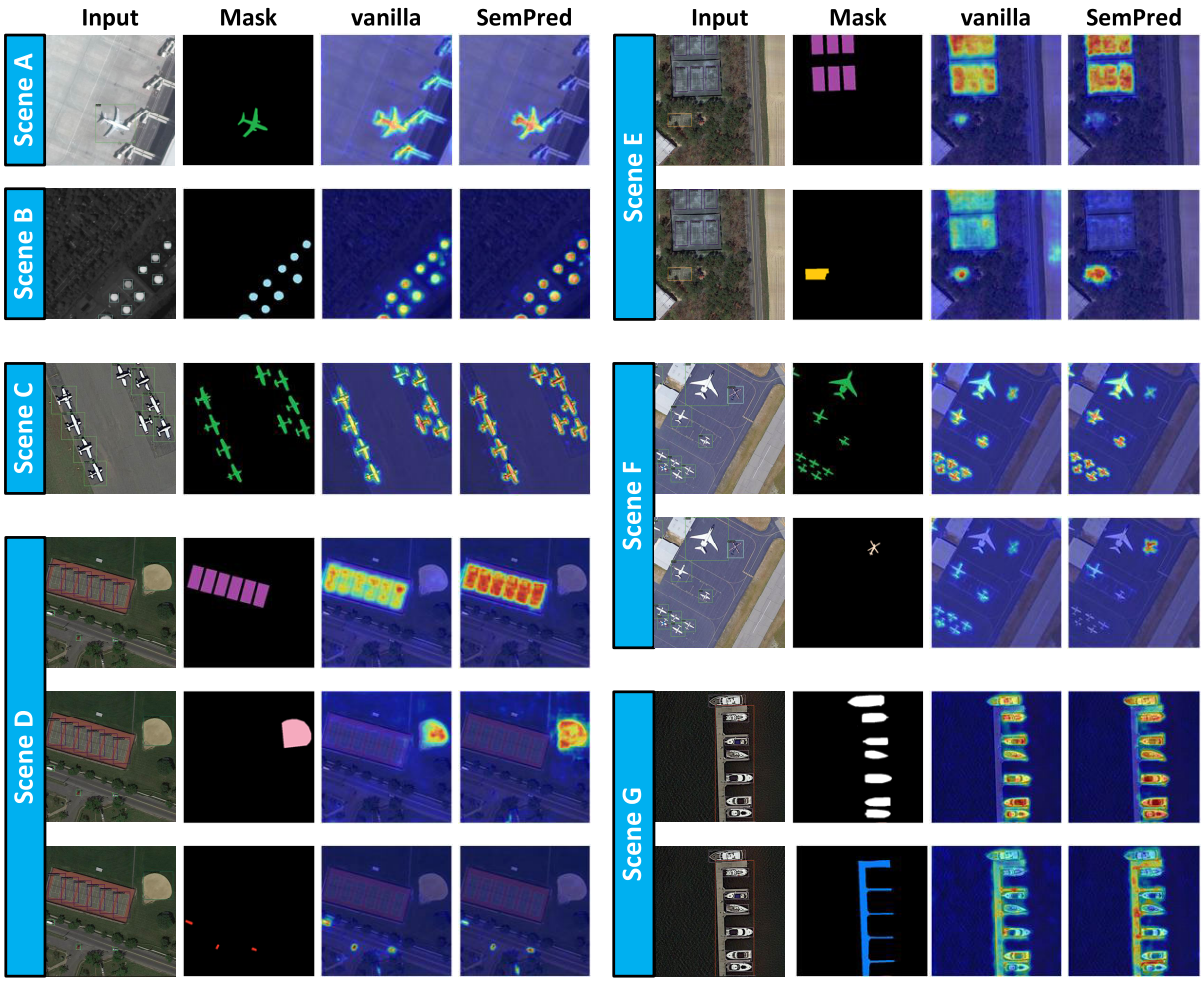


Fig. 6. Heatmaps of specific response layers produced by our SemPred module with and without performing SFG.

TABLE III

COMPARISON OF THE PSEUDO-BOX QUALITY AND DETECTION PERFORMANCE ACHIEVED BY OUR PLUG WITH DIFFERENT SEMPRED MODULES. NOTE THAT THE VANILLA AND SEMPRED MODULES REPRESENT THE METHOD WITHOUT AND WITH PERFORMING SFG, RESPECTIVELY

semantic prediction module	Pseudo box quality				Detection performance			
	$mIoU$	$mIoU_s$	$mIoU_m$	$mIoU_l$	mAP_{50}	mAP_s	mAP_m	mAP_l
vanilla	0.497	0.494	0.512	0.457	0.356	0.292	0.401	0.215
SemPred	0.549	0.539	0.576	0.533	0.427	0.329	0.474	0.338

response maps produced by our PLUG and its variants (vanilla and SemPred). We can draw the following conclusions from Fig. 6.

1) The SFG scheme can improve the recognition capability of our PLUG on confusing backgrounds. As shown in Scene A, the plane (PL) and the boarding bridges are similar in color space. With the guidance of sparse features, our PLUG can better distinguish objects from the background.
2) The SFG scheme can improve the recognition capability of our PLUG on dense objects. For densely packed objects of the same category (e.g., Scenes B and C), some objects are weakly activated when SFG is not performed. In contrast, by performing SFG, the features of each object can be enhanced, and the intraclass instance recognition performance is improved. Besides, SFG can also improve the recognition capability of our PLUG on densely packed objects of different categories [e.g., the ships (SH) and harbor (HB) in Scene G].
3) The SFG scheme can enhance the capability of our PLUG to distinguish objects in different categories but with similar appearances. As shown in Scene E, the tennis court (TC) and basketball court (BC) have similar appearances, and our PLUG without SFG cannot distinguish them and produces a falsely mixed response. Since category-aware meta-features are used to aggregate the extracted features, the enhanced features have stronger category characteristics. Consequently, our PLUG with SFG can effectively handle this mixed response issue and can well distinguish similar objects.

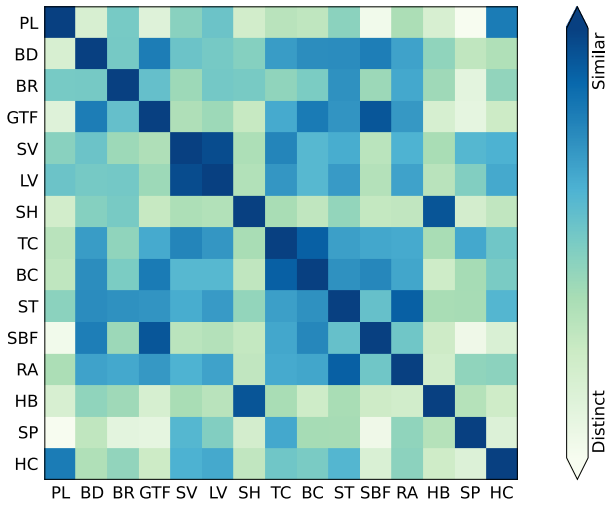


Fig. 7. Cosine similarities between different pairs of representations in meta-features. Here, darker colors indicate larger values (i.e., higher similarity).

Cross-Category Correlation of Meta-Features: Meta-features are the general representation of objects in different categories. Here, we visualize the cosine similarity map between each pair of meta-features to investigate their correlation. As shown in Fig. 7, apart from the elements on the diagonal, there are still some pairs of meta-features [e.g., large vehicle (LV) versus small vehicle (SV), plane (PL) versus helicopter (HC), basketball court (BC) versus tennis court (TC)] highly correlated due to the similar appearance of the objects. This observation is consistent with the visualization results in Fig. 6 and can demonstrate the effectiveness of the usage of meta-features.

3) Effectiveness of the Edge-Aware Neighbor Cost: In this section, we validate the effectiveness of the edge-aware neighbor cost in the ILG module. Fig. 8 shows the likelihood maps $P_{\text{map}} = 1 - C_{\text{map}}$ with and without using edge-aware neighbor cost on an example scene, where the values represent the likelihood of a pixel belonging to a specific instance. It can be observed that densely packed adjacent instances cannot be well distinguished without using edge-aware neighbor cost. That is because the semantic-aware neighbor cost encourages the labeled points to diffuse to the adjacent semantic-similar areas and tends to consider the densely packed objects as a single instance. When the edge-aware neighbor cost is introduced, the diffusion of labeled points can stop at the boundaries, and these densely packed objects can be better distinguished.

Note that, the value of λ in (2) should be properly set to ensure preferable growth from point labels. We compare the quality of pseudo-boxes and the detection performance with respect to different λ values. As shown in Table IV, when λ is set to 0.5, our PLUG can generate pseudo-boxes of the highest quality, and our PLUG-Det can achieve the best detection performance. Consequently, we set λ to 0.5 to balance the semantic-aware and edge-aware neighbor cost.

4) Effectiveness of Losses: In this section, we conduct ablation studies to validate the effectiveness of the proposed losses. As shown in Table V, our PLUG can only achieve an mIoU of 0.318 when the positive loss is used only. That is because the background cannot be considered in the training process and thus degrades the recognition capability of our PLUG to distinguish objects and backgrounds. When the negative loss is introduced, both the quality of pseudo-boxes and the detection performance are significantly improved. Moreover, applying the color prior loss can further introduce a 0.051 improvement of mIoU and a 0.006 improvement of mAP_{50}. The experimental results demonstrate the effectiveness of the proposed losses.

TABLE IV

COMPARISON OF THE PSEUDO-BOX QUALITY AND DETECTION PERFORMANCE ACHIEVED BY OUR PLUG WITH DIFFERENT λ VALUES. BEST RESULTS ARE IN BOLD FACES

λ	Pseudo box quality				Detection performance			
	mIoU	$mIoU_s$	$mIoU_m$	$mIoU_l$	mAP_{50}	mAP_s	mAP_m	mAP_l
0	0.497	0.473	0.552	0.548	0.405	0.305	0.461	0.340
0.5	**0.549**	0.539	**0.576**	**0.533**	**0.427**	**0.329**	0.474	**0.338**
1.0	0.547	**0.541**	0.567	0.528	0.425	0.327	0.467	0.319
1.5	0.542	0.536	0.559	0.523	0.426	0.322	**0.475**	0.330
2.0	0.517	0.389	0.547	0.552	0.422	0.328	0.473	0.335

TABLE V

COMPARISON OF THE PSEUDO-BOX QUALITY AND DETECTION PERFORMANCE ACHIEVED BY OUR PLUG WITH DIFFERENT LOSSES

Loss			*mIoU*	mAP_{50}
positive	negative	color prior		
✓			0.318	0.175
✓	✓		0.498	0.421
✓	✓	✓	**0.549**	**0.427**

D. Analyses of the Selecting Strategy of Point Labels

In the preceding experiments, point labels were randomly selected from object masks. How will the locations of the selected points affect the performance? In this section, we implement three kinds of point labels and conduct experiments to analyze their impacts on the quality of generated pseudo-boxes and downstream detection performance. Specifically, we adopt three different labeling strategies, that is, selecting the point in the center, selecting the point on the corner, and randomly selecting a point on the mask. Note that, since there is no clear definition of the corners of objects, we just selected the point (on the mask) that is farthest from the center point as its "corner" label. Objects with different point labels are shown in Fig. 9.

Table VI shows the quality of pseudo-boxes and the detection performance of our method with different point labels. It can be observed that our PLUG with center point labels can achieve the most superior results, which are 0.553 in terms of mIoU and 0.438 in terms of mAP_{50}. Besides, when the randomly selected points are used, the performance is slightly decreased (0.549 and 0.427 in terms of mIoU and mAP_{50}, respectively). Moreover, the corner labels result in a larger

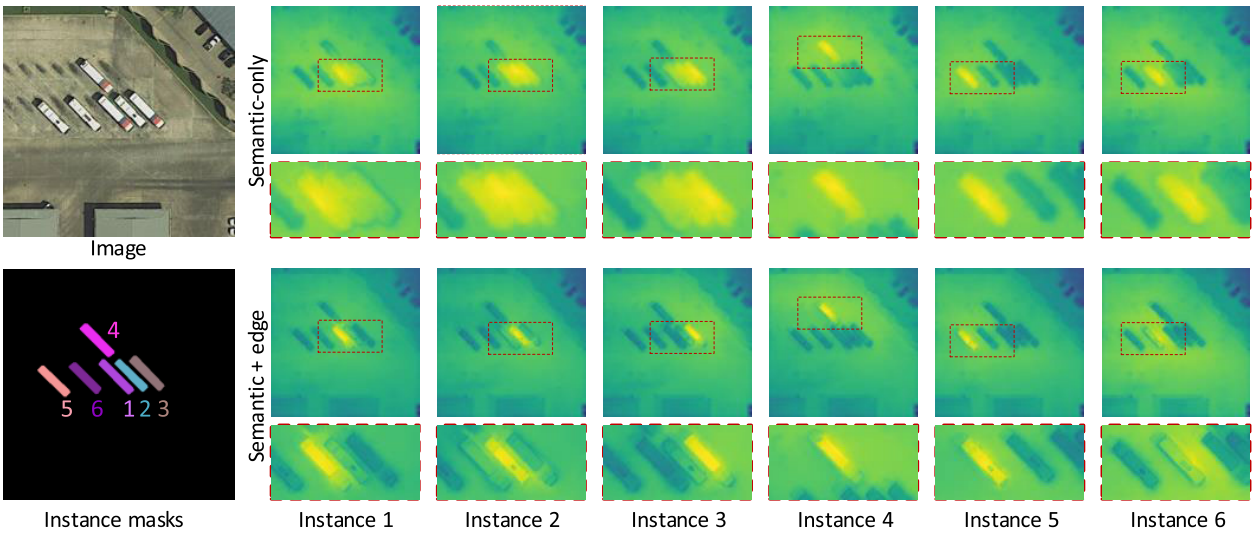


Fig. 8. Likelihood maps generated by the ILG module with and without using the edge-aware neighbor cost. Note that, we visualize $P_{\text{map}} = 1 - C_{\text{map}}$ for better visual analyses, where C_{map} is the cost map for each labeled point, and the values on the cost map can represent the costs from each pixel to the labeled point. Consequently, $C_{\text{map}} \in H \times W \times N$, where H and W are the height and width of images and N is the number of objects in the image. Based on C_{map}, P_{map} can represent the likelihood of each pixel belonging to a specific instance, and thus can more intuitively show the diffusion of labeled points.

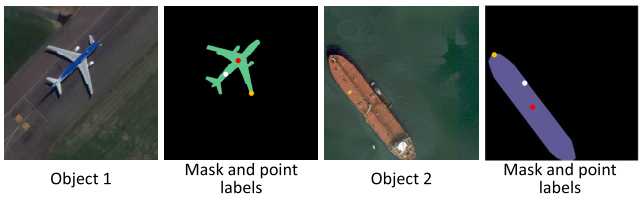


Fig. 9. Objects with point labels under different point selection strategies. The red points are the centers of masks, the yellow points are the corners of masks, and the white points are the randomly selected points on masks.

TABLE VI

COMPARISON OF THE PSEUDO-BOX QUALITY AND DETECTION PERFORMANCE ACHIEVED BY OUR PLUG WITH DIFFERENT POINT LABEL SELECTION STRATEGIES

Selection	Pseudo box quality				Detection performance			
Strategy	*mIoU*	$mIoU_s$	$mIoU_m$	$mIoU_l$	mAP_{50}	mAP_s	mAP_m	mAP_l
corner	0.518	0.488	0.586	0.589	0.406	0.306	0.464	0.427
center	0.553	0.520	0.629	0.606	0.438	0.316	0.493	0.504
random	0.549	0.539	0.576	0.533	0.427	0.329	0.474	0.338

degree of performance degradation, in which mIoU and mAP_{50} are decreased to 0.518 and 0.406, respectively.

It is worth noting that the performance of our method with corner point labels is inferior to that with center and random point labels. That is because the edge-aware neighbor cost used in the ILG module hinders the pixel diffusion of corner points. Specifically, the edge-aware neighbor cost is utilized to help stop the diffusion of labeled points at boundaries and thus prevent the labeled points from spreading toward the background areas (see Section IV-C3). However, since the corner points are located on the boundaries of objects, the edge-aware cost may hinder the diffusion of the labeled points to the internal area of the object, as their paths pass through the edges. For example, as shown in P_{map} of the instance 6 in Fig. 8, the ILG module can recognize its correct regions with the semantic-aware cost only. However, when the edge-aware cost is introduced, the labeled points can only be diffused to background areas.

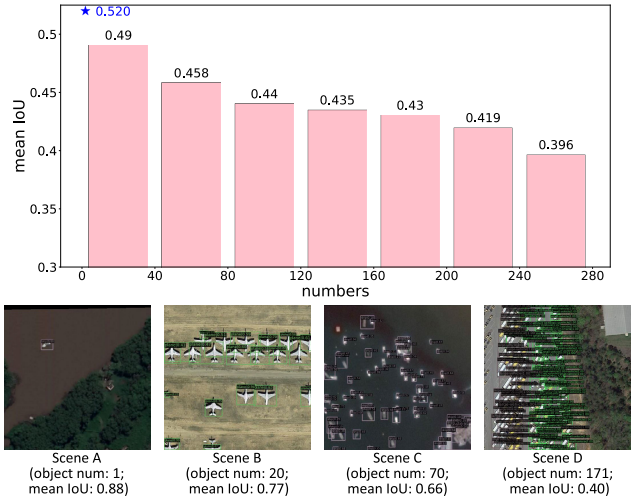


Fig. 10. IoU of generated pseudo-boxes in images with different numbers. Here, four exampled scenes are shown for visualization. Note that the blue star indicates that the mean IoU of pseudo-boxes is 0.520 in images with a single object.

E. Extension to Rotated Object Detection

In our method, the ILG module utilizes semantic and edge information to assign pixels to its most likely object or background and uses the circumscribed rectangle of assigned pixels as pseudo boxes. Therefore, by further transforming the circumscribed rectangle to the one with the minimum area, our method can be easily extended to the task of rotated object detection. We conduct experiments to validate the effectiveness of our method on rotated object detection. Specifically, we use the modified PLUG to generate rotated pseudo-boxes and use ROITrans [26] as the downstream rotated detector to

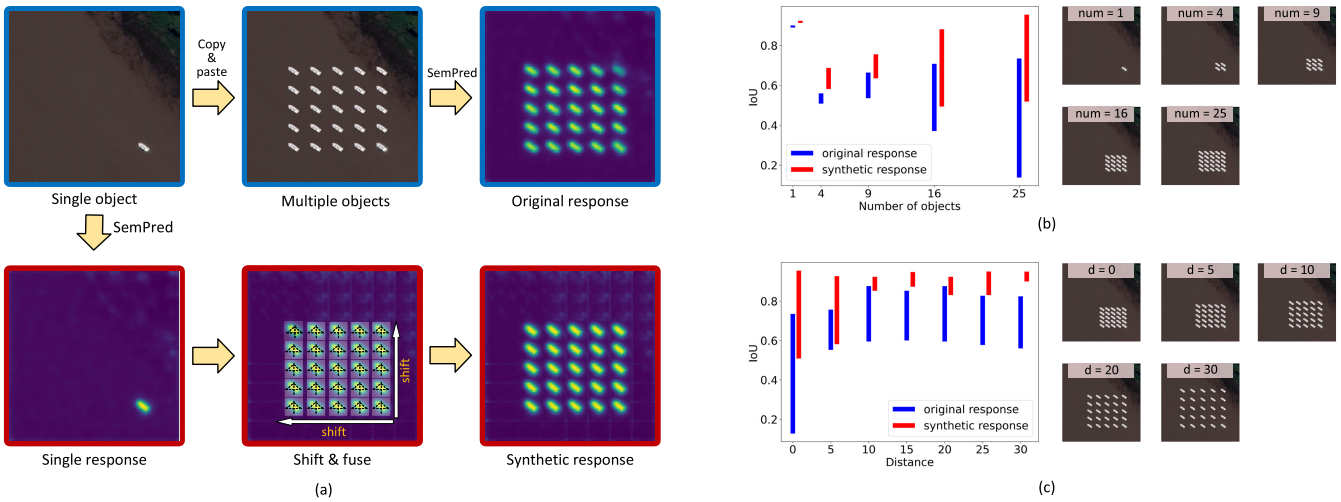


Fig. 11. Illustrations and analyses of the influence of densely packed objects on the quality of generated pseudo-boxes. (a) Illustration of the generation process of the two kinds of semantic response maps. Note that the synthetic response is obtained by shifting the response map of a single object with specific offsets, and then fusing the shifted response. (b) IoU score ranges of the pseudo-boxes generated from the two kinds of semantic response maps in (a) with varying object numbers. (c) IoU score ranges of the pseudo-boxes generated from the two kinds of semantic response maps in (a) with varying object distances (in pixels).

TABLE VII
COMPARISON OF THE DETECTION PERFORMANCE ACHIEVED BY ROITRANS AND PLUG-ROITRANS

Method	Supervision	Backbone	Categories															mAP_{50}
			PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HB	SP	HC	
PLUG-ROITrans	Point	ResNet50	0.088	0.503	0.302	0.292	0.248	0.661	0.368	0.806	0.285	0.502	0.400	0.365	0.259	0.176	0.009	0.351
ROITrans [26]	Rotated Box	ResNet50	0.798	0.671	0.500	0.736	0.713	0.851	0.885	0.906	0.551	0.693	0.620	0.651	0.676	0.578	0.366	0.680

develop PLUG-ROITrans. The experimental results of our PLUG-ROITrans (under single-point supervision) and the original ROITrans (under ground-truth rotated box supervision) are shown in Table VII. It can be observed that our PLUG-ROITrans can achieve 0.351 in terms of mAP_{50}, which is 51.6% of the performance of fully supervised ROITrans. The results demonstrate the preliminary effectiveness of our method in pointly supervised rotated object detection in RS images.

F. Further Analyses on Dense Objects

As mentioned in Section I, dense objects introduce challenges to discriminative feature extraction and thus affect the quality of generated pseudo-boxes. In this section, we conduct a series of experiments to analyze the influence of dense objects. First, we coarsely suppose that the density of objects is positively related to their numbers in an image patch (with the same area). Then, we split the DOTA dataset into several subsets containing different numbers of objects and quantitatively evaluate the quality of generated pseudo-boxes with respect to the object density. Note that, we do not perform SFG in our PLUG to better demonstrate the challenges introduced by dense objects. As shown in Fig. 10, the quality of generated boxes degrades as the number of objects (i.e., density) increases. The examples in Scenes A to D qualitatively illustrate the quality degradation of pseudo-boxes with dense objects.

Second, considering that the number, adjacent distance, and appearance of objects are the three key factors that influence the quality of pseudo-boxes, we design specific experiments to quantitatively investigate the impact of the first two factors by keeping the object appearance unchanged. Specifically, we use the "copy-and-paste" strategy [see the subfigures with blue boxes in Fig. 11(a)] to generate multiple identical objects with controllable density. As shown in Fig. 11(b) and (c), the quality of generated boxes degrades as the object density increases.

Finally, we keep the density of the semantic response maps unchanged and investigate the influence of densely packed objects on the discriminative feature extraction. Specifically, we shift and fuse the single-object response to synthesize a pseudo-dense-object response map. In this way, we build a control group with identical object density in the response maps but different feature representations in the feature extraction module. As shown in Fig. 11(b) and (c), the mIoU scores of the pseudo-boxes generated from the control group are significantly higher than those obtained from the images with dense objects. The experimental results clearly validate that densely packed objects in RS images can hinder the discriminative feature extraction and thus degrade the quality of pseudo boxes. With our SFG scheme, the mIoUs of generated pseudo-labels in different density intervals are increased. The qualitative results demonstrate the effectiveness of our method in handling densely packed objects.

G. Extension to Instance Segmentation

Since the ILG module in our PLUG produces instance labels for each object from their point annotation, our PLUG can be easily extended to PSIS. Specifically, we concatenated

TABLE VIII
QUANTITATIVE RESULTS ACHIEVED BY DIFFERENT INSTANCE SEGMENTATION METHODS ON THE DOTA DATASET

model	supervision	object detection				instance segmentation			
		mAP_{50}	mAP_s	mAP_m	mAP_l	mAP_{50}	mAP_s	mAP_m	mAP_l
Mask-RCNN [87]	*Mask*	0.659	0.535	0.670	0.697	0.623	0.480	0.662	0.682
BoxInst [91]	*Box*	0.643	0.535	0.633	0.647	0.503	0.371	0.543	0.582
PLUG-Seg (ours)	*Point*	0.435	0.335	0.481	0.348	0.406	0.278	0.491	0.340

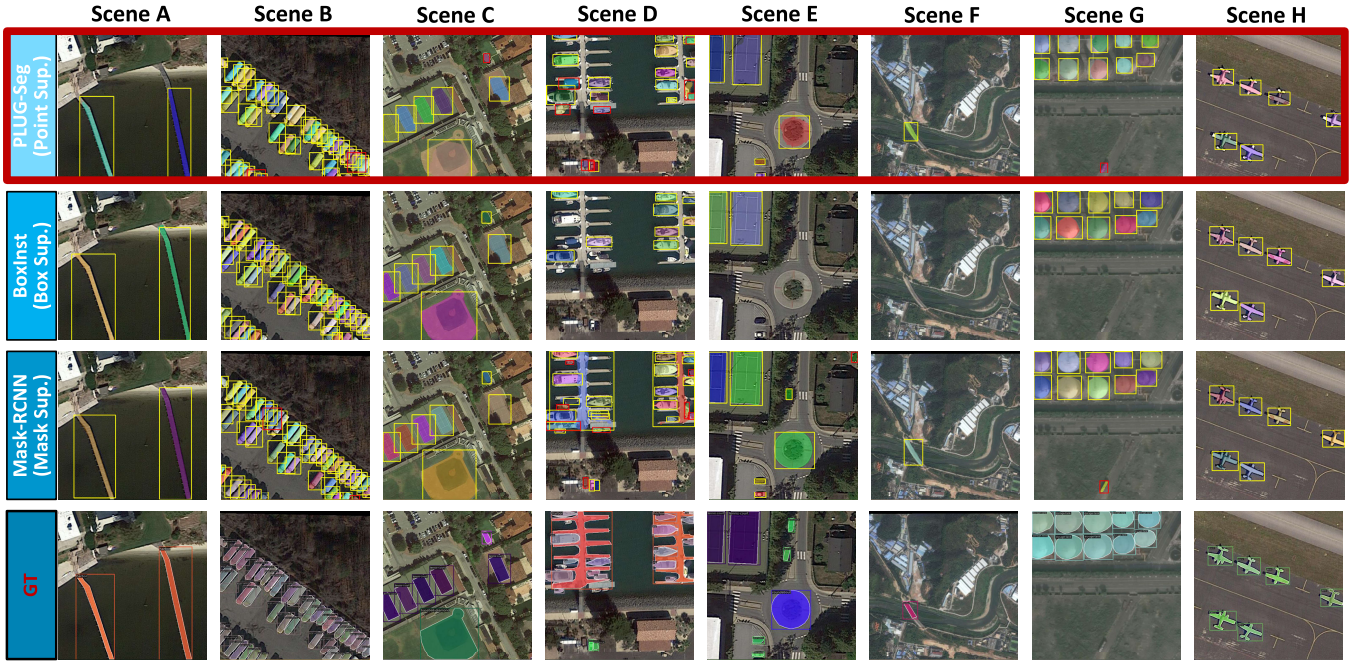


Fig. 12. Qualitative results achieved by different instance segmentation methods on the DOTA dataset. The correctly detected results are marked by yellow boxes, and the falsely detected results are marked by red boxes. Gradually darker colors represent stronger supervision. The predicted instance masks are randomly colored.

our PLUG with Mask-RCNN and developed a PLUG-Seg network to achieve PSIS in RS images. Besides, we used the ground-truth mask labels in the iSAID dataset [88] and adopted the mask-level mAP_{50}, mAP_s, mAP_m, and mAP_l as quantitative metrics for performance evaluation. We compare our PLUG-Seg with BoxInst [91] and Mask-RCNN [87], which use box-level and mask-level supervision for instance segmentation, respectively. We also followed these two methods [87], [91] to evaluate the performance of object detection and instance segmentation simultaneously. The experimental results are shown in Table VIII and Fig. 12.

It can be observed from Table VIII that our PLUG-Seg can achieve an mAP_{50} of 0.435 for object detection and an mAP_{50} of 0.406 for instance segmentation. With single-point annotation for each instance, our PLUG-Seg can achieve 68%/81% and 66%/65% accuracy in object detection/instance segmentation when compared to box-level (i.e., BoxInst [91]) and mask-level (i.e., Mask-RCNN [87]) supervised methods, respectively. The qualitative results in Fig. 12 also demonstrate the promising performance of our PLUG-Seg. It is worth noting that our PLUG-Seg can achieve better performance than BoxInst [91] on scenes with complex backgrounds (e.g., the roundabout and small vehicles in Scene E and the bridge in Scene F). These experimental results demonstrate that single-point annotation can provide sufficient supervision for instance segmentation.

V. CONCLUSION

In this article, we proposed a method to learn RSOD with single-point supervision. In our method, a PLUG is designed to generate pseudo-boxes from point labels. We also handle the dense object issue in RS images by designing a sparse feature-guided SemPred module. Experimental results validate the effectiveness and superiority of our method. In the future, we will further extend our method to generate rotated pseudo-boxes from single-point labels and investigate more stable and efficient pseudo-label generation schemes. We hope that our study can draw attention to the research of single-point supervised RSOD.

REFERENCES

[1] L. Hou, K. Lu, and J. Xue, "Refined one-stage oriented object detection method for remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 1545–1558, 2022.

[2] 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.

[3] Z. Huang, W. Li, X.-G. Xia, and R. Tao, "A general Gaussian heatmap label assignment for arbitrary-oriented object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 1895–1910, 2022.

[4] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.

[5] B. Liu, C. Xu, Z. Cui, and J. Yang, "Progressive context-dependent inference for object detection in remote sensing imagery," *IEEE Trans. Image Process.*, vol. 32, pp. 580–590, 2023.

[6] W. Li, W. Wei, and L. Zhang, "GSDet: Object detection in aerial images based on scale reasoning," *IEEE Trans. Image Process.*, vol. 30, pp. 4599–4609, 2021.

[7] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511.

[8] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, vol. 2, 2017, pp. 324–331.

[9] 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.

[10] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

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

[12] 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.

[13] 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.

[14] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8291–8299.

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

[16] J. Xie, C. Luo, X. Zhu, Z. Jin, W. Lu, and L. Shen, "Online refinement of low-level feature based activation map for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 132–141.

[17] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari, "Training object class detectors with click supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 180–189.

[18] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, A. G. Schwing, and J. Kautz, "UFO^2: A unified framework towards omni-supervised object detection," in *Proc. ECCV*. Glasgow, U.K.: Springer, 2020, pp. 288–313.

[19] P. Chen et al., "Point-to-box network for accurate object detection via single point supervision," in *Proc. ECCV*. Tel Aviv-Yafo, Israel: Springer, 2022, pp. 51–67.

[20] X. Yu et al., "Object localization under single coarse point supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4868–4877.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[22] 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.

[23] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8231–8240.

[24] J. Ma et al., "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[25] Y. Jiang et al., "R2 CNN: Rotational region CNN for arbitrarily-oriented scene text detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3610–3615.

[26] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2844–2853.

[27] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3500–3509.

[28] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, Nov. 2020.

[29] H. Wei, Y. Zhang, B. Wang, Y. Yang, H. Li, and H. Wang, "X-LineNet: Detecting aircraft in remote sensing images by a pair of intersecting line segments," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1645–1659, Feb. 2021.

[30] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2149–2158.

[31] P. Zhao, Z. Qu, Y. Bu, W. Tan, and Q. Guan, "PolarDet: A fast, more precise detector for rotated target in aerial images," *Int. J. Remote Sens.*, vol. 42, no. 15, pp. 5831–5861, Aug. 2021.

[32] P. Dai, S. Yao, Z. Li, S. Zhang, and X. Cao, "ACE: Anchor-free corner evolution for real-time arbitrarily-oriented object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 4076–4089, 2022.

[33] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. ICML*, 2021, pp. 11830–11841.

[34] X. Yang et al., "The KFIoU loss for rotated object detection," 2022, *arXiv:2201.12558*.

[35] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2785–2794.

[36] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI*, 2021, vol. 35, no. 4, pp. 3163–3171.

[37] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," in *Proc. AAAI Conf. Arti. Intell.*, May 2021, vol. 35, no. 3, pp. 2458–2466.

[38] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. ECCV*. Glasgow, U.K.: Springer, 2020, pp. 677–694.

[39] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15814–15824.

[40] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.

[41] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 1968–1979, 2022.

[42] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8310–8319.

[43] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 737–746.

[44] K. Fu, Z. Chang, Y. Zhang, and X. Sun, "Point-based estimator for arbitrary-oriented object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4370–4387, May 2021.

[45] E. Liu, Y. Zheng, B. Pan, X. Xu, and Z. Shi, "DCL-Net: Augmenting the capability of classification and localization for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7933–7944, Sep. 2021.

[46] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "RFLA: Gaussian receptive field based label assignment for tiny object detection," in *Proc. ECCV*. Tel Aviv-Yafo, Israel: Springer, 2022, pp. 526–543.

[47] Y. Zhu, J. Du, and X. Wu, "Adaptive period embedding for representing oriented objects in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7247–7257, Oct. 2020.

[48] S. Liu, L. Zhang, H. Lu, and Y. He, "Center-boundary dual attention for oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603914.

[49] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*. Glasgow, U.K.: Springer, 2020, pp. 213–229.

[50] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[51] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query DeNoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13609–13617.

[52] L. Dai, H. Liu, H. Tang, Z. Wu, and P. Song, "AO2-DETR: Arbitrary-oriented object detection transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2342–2356, May 2023.

[53] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 607315.

[54] C. Fasana, S. Pasini, F. Milani, and P. Fraternali, "Weakly supervised object detection for remote sensing images: A survey," *Remote Sens.*, vol. 14, no. 21, p. 5362, Oct. 2022.

[55] G. Cheng, J. Yang, D. Gao, L. Guo, and J. Han, "High-quality proposals for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 5794–5804, 2020.

[56] 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.

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

[58] F. Shao et al., "Deep learning for weakly-supervised object detection and localization: A survey," *Neurocomputing*, vol. 496, pp. 192–207, Jul. 2022.

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

[60] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. ECCV*. Zürich, Switzerland: Springer, 2014, pp. 391–405.

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

[62] 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.

[63] 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.

[64] X. Qian, C. Li, W. Wang, X. Yao, and G. Cheng, "Semantic segmentation guided pseudo label mining and instance re-detection for weakly supervised object detection in remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 119, May 2023, Art. no. 103301.

[65] B. Wang, Y. Zhao, and X. Li, "Multiple instance graph learning for weakly supervised remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5613112.

[66] X. Feng, X. Yao, G. Cheng, and J. Han, "Weakly supervised rotation-invariant aerial object detection network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14126–14135.

[67] G. Wang, X. Zhang, Z. Peng, X. Jia, X. Tang, and L. Jiao, "MOL: Towards accurate weakly supervised remote sensing object detection via multi-view nOisy learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 196, pp. 457–470, Feb. 2023.

[68] 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.

[69] X. Qian et al., "Incorporating the completeness and difficulty of proposals into weakly supervised object detection in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1902–1911, 2022.

[70] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 549–565.

[71] R. Qian, Y. Wei, H. Shi, J. Li, J. Liu, and T. Huang, "Weakly supervised scene parsing with point-based distance metric learning," in *Proc. AAAI*, 2019, vol. 33, no. 1, pp. 8843–8850.

[72] L. Wu, L. Fang, J. Yue, B. Zhang, P. Ghamisi, and M. He, "Deep bilateral filtering network for point-supervised semantic segmentation in remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 7419–7434, 2022.

[73] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt, "Proposal-based instance segmentation with point supervision," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2126–2130.

[74] B. Cheng, O. Parkhi, and A. Kirillov, "Pointly-supervised instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2607–2616.

[75] M. Liao, Z. Guo, Y. Wang, P. Yuan, B. Feng, and F. Wan, "AttentionShift: Iteratively estimated part-based attention map for pointly supervised instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19519–19528.

[76] J. Fan, Z. Zhang, and T. Tan, "Pointly-supervised panoptic segmentation," in *Proc. ECCV*. Tel Aviv-Yafo, Israel: Springer, 2022, pp. 319–336.

[77] J. Ribera, D. Guera, Y. Chen, and E. J. Delp, "Locating objects without bounding boxes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6472–6482.

[78] Q. Song et al., "Rethinking counting and localization in crowds: A purely point-based framework," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3345–3354.

[79] X. Ying et al., "Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15528–15538.

[80] B. Li et al., "Monte Carlo linear clustering with single-point supervision is enough for infrared small target detection," 2023, *arXiv:2304.04442*.

[81] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[82] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[83] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[84] R. O. Duda et al., *Pattern Classification and Scene Analysis*, vol. 3. New York, NY, USA: Wiley, 1973.

[85] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*.

[86] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.

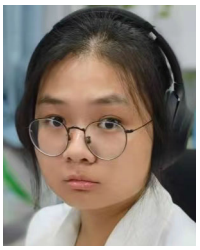
[87] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[88] S. W. Zamir et al., "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. CVPRW*, 2019, pp. 28–37.

[89] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.

[90] H. Robbins and S. Monro, "A stochastic approximation method," *AMS*, vol. 22, no. 3, pp. 400–407, 1951.

[91] Z. Tian, C. Shen, X. Wang, and H. Chen, "BoxInst: High-performance instance segmentation with box annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5439–5448.



Shitian He received the B.E. degree in electronic and information engineering from Beijing Jiaotong University (BJTU), Beijing, China, in 2019, and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2021, where she is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology.

Her research interests include object detection and tracking.

Huanxin Zou was born in Meizhou, Guangdong, China, in 1973. He received the B.S. degree in information engineering and the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 1995, 2000, and 2003, respectively.

He held a visiting position with the Department of Computing Science, University of Alberta, Edmonton, AB, Canada, from March 2015 to September 2015. He is currently a Professor with the College of Electronic Science and Technology, National University of Defense Technology. His main research interests include computer vision, deep learning, pattern recognition, and remote-sensing image processing and interpretation.

Yingqian Wang (Member, IEEE) received the B.E. degree in electrical engineering from Shandong University (SDU), Jinan, China, in 2016, and the M.E. and D.E. degrees in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018 and 2023, respectively.

He is currently a Lecturer with the College of Electronic Science and Technology, NUDT. His research interests include low-level vision, particularly in light field imaging and image super-resolution.

Boyang Li received the B.E. degree in mechanical design manufacture and automation from Tianjin University, Tianjin, China, in 2017, and the M.S. degree in biomedical engineering from the National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing, China, in 2020. He is currently pursuing the Ph.D. degree in information and communication engineering with the National University of Defense Technology (NUDT), Changsha, China.

His research interests include infrared small target detection, weakly supervised semantic segmentation, and deep learning.

Xu Cao received the B.E. degree in information engineering and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018 and 2021, respectively, where he is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology.

His research interests include object detection and multisource information fusion interpretation.

Ning Jing received the Ph.D. degree in computer science from the National University of Defense Technology, Changsha, China, in 1990.

He is currently a Professor with the National University of Defense Technology. His research interests include the management and analysis of big data and spatial information systems.