Enhancing Prospective Consistency for Semi-Supervised Object Detection in Remote Sensing Images

Jinhao Shen, Cong Zhang, Yuan Yuan, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Deep learning-based object detection has recently played a vital role in both computer vision and Earth observation communities. However, the performance of modern object detectors is highly limited by the quantity and quality of manually labeled training samples. Furthermore, compared to object detection in natural scenes, Remote Sensing Object Detection (RSOD) faces two specific critical challenges. 1) Densely arranged instances: geospatial objects tend to be densely packed in remote sensing scenarios. 2) Large variations in object scale: the wide field of the bird's eye view leads to dramatic variations in object scale across various categories. The above issues bring significant difficulties to attaining manual annotations for deep learning-based RSOD. To this end, in this paper, we turn our attention from fully-supervised RSOD to semi-supervised RSOD, and propose a novel framework based on the teacher-student paradigm, namely Prospective Consistent Teacher (PCT), which includes three crucial components, *i.e.*, Weighted Dense-Proposal Learning (WDPL), Mean-Consistency-based Proposal Pruning (MCPP), and EM-based Fitting Policy (EFP). Specifically, WDPL re-weights the dense proposals with box confidences, while MCPP ranks the student proposals with consistency analysis to select discriminative and consistent boxes. EFP can automatically set thresholds for pseudo labels and improve the consistent information of the teacher network. Extensive experimental results on two challenging public datasets, *i.e.*, DOTA and DIOR, have demonstrated the reduced reliance of our proposed method on large amounts of labeled data for the task of RSOD.**

***Index Terms*—remote sensing images, semi-supervised object detection, teacher-student network.**

I. INTRODUCTION

AS a significant part of computer vision, object detection is one of the most crucial and essential tasks. Driven by sufficient labeled data, modern object detection methods have superior performance so that they have facilitated many practical applications, such as resource exploration and intelligent surveillance systems [1], [2], [3], [4], [5], [6]. In the past decade, Earth Vision technology [7] observes the surface of the Earth with an aerial view and has numerous real-world applications, including satellite monitoring and geoscience interpretation [8], [9]. Remote Sensing Object Detection (RSOD) is a fundamental assignment for these applications, so it has been widely studied, [10], [11], [12]. RSOD aims to recognize the category and localization of densely distributed instances within remote sensing images that exhibit complex background characteristics.

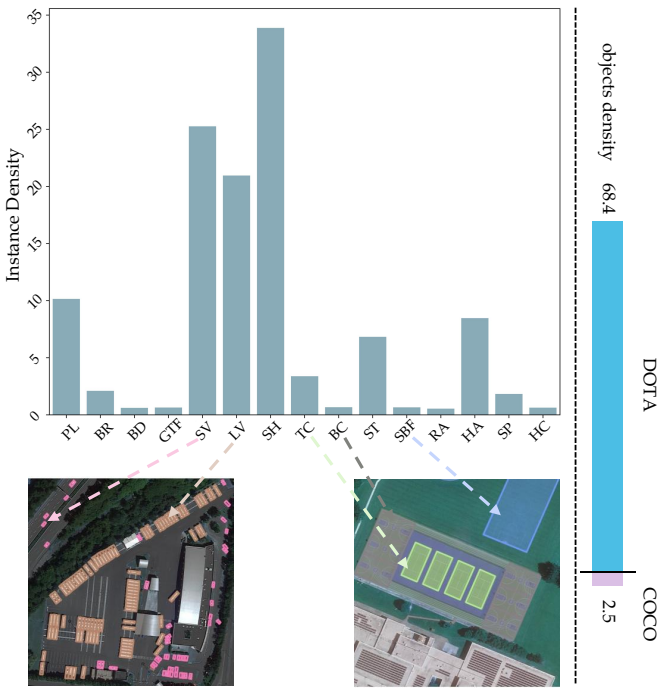

Fig. 1: Analysis of remote sensing dataset: DOTA. The histogram illustrates the object density of various categories in DOTA. The right part illustrates a comparison of the average object density between DOTA and MS-COCO.

Nevertheless, the expensive manual annotation process makes it tough to generate sufficient and high-quality labels. It encourages many researchers to investigate unsupervised data without substantially affecting network performance [13], [14], [15]. Fig. 1 illustrates the statistical differences between the remote sensing and natural scene datasets, with DOTA [16] and MS-COCO [17] serving as examples. Obviously, a single remote sensing picture typically comprises dozens of targets of various sizes. These crowded and multi-scale objects can lead to more complicated and controversial annotations. Semi-supervised object detection (SSOD) aims to combine a small

Jinhao Shen, Yuan Yuan, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China. (e-mail: jinhaoshen00@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com)

Cong Zhang is with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. (e-mail: congclarence.zhang@connect.polyu.hk)

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, 61825603, National Key R&D Program of China 2020YFB2103902.

Corresponding author is Qi Wang.

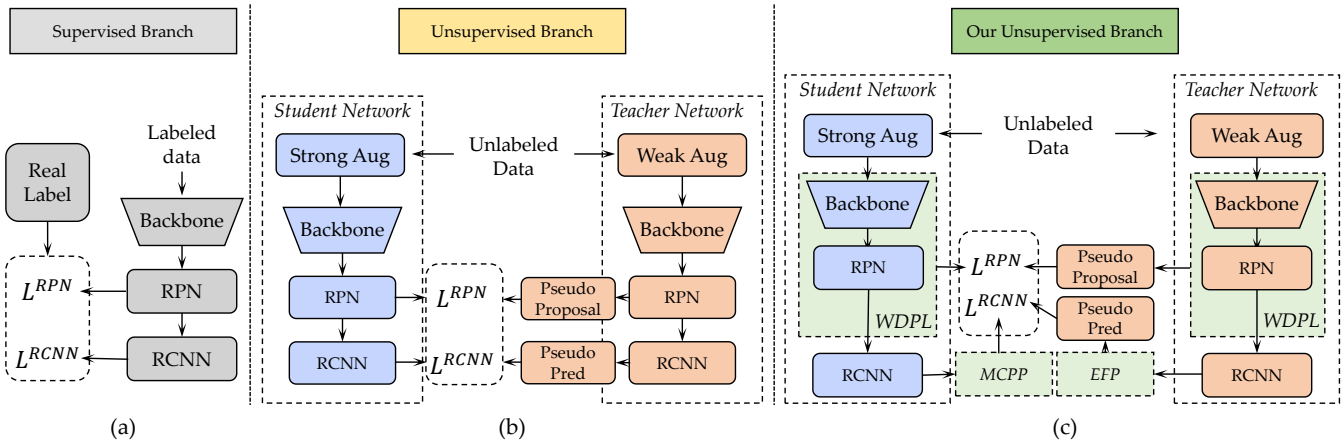

Fig. 2: The concept map of Teacher-Student Framework. (a) and (b) depict the supervised and unsupervised branches in the vanilla teacher-student frameworks. (c) depicts the unsupervised branch of our proposed method.

set of labeled images with a large set of unlabeled images to enhance the localization accuracy in a specific class of objects. It is proven to be effective and robust for extracting unsupervised information in natural images [18], [19], [20], [21].

In practice, generic SSOD methods employ the vanilla Teacher-Student Framework (TSF) [22], [23], [24], [25], as illustrated in (a) and (b) of Fig. 2. It contains two detectors: student and teacher networks. Additionally, the training schedule is split into dual parallel branches on the basis of labels, referred to as the supervised branch and the unsupervised branch. In the supervised branch, labeled data is trained by the student in the same way as the general object detection. In contrast, the teacher detector in the unsupervised branch computes the features from pre-processed images. The predicted boxes produced by the teacher network are considered pseudo labels. Meanwhile, the student detector processes the same images with various augmentations and then learns consistent information from pseudo labels. Lastly, the teacher network updates the model weights from the student via Exponential Moving Average (EMA) [26] after each iteration. However, classic natural scenes are sparsely arranged and with small-scale variations, which causes the generic SSOD methods do not match remote sensing images. Therefore, there is a considerable research need for remote sensing SSOD.

The teacher and student networks usually comprise ambiguous and damaged information in various modules, namely inconsistency. It severely restricts the paradigm's capacity to generalize and understand. Specifically, the major inconsistent obstacles in remote sensing semi-supervised object detection tasks are introduced as follows:

1) The first issue is the inconsistency by high instance density. Fig. 1 demonstrates the primary statistics of DOTA, including its dense-instance characteristic. The student and teacher networks contain a higher degree of conflicting information especially when extracting low- and high-level features from dense objects. Generic SSOD methods will produce more misaligned features and redundant detection boxes, thereby constraining the accuracy of object localization.
2) The second issue is the inconsistency by multi-scale variation. In Fig. 1, objects scale in remote sensing images can vary sharply. It brings noisy and complex inconsistency to networks between the student and teacher detectors. When generic SSOD methods are adopted to restrict labeled data during training, there are several targets at different scales been falsely identified or missed.

Most existing SSOD methods are developed for natural scene images, yet overlook the inconsistency of dense and multi-scale targets. As a result, semi-supervised detection in remote sensing still encounters significant challenges. Hence, it is imperative to devise a novel teacher-student network to address the inconsistency arising from remote sensing scenarios.

To tackle the first issue, an efficient approach called Weighted Dense-Proposal Learning (WDPL) is presented. It involves dual parallel pathways for feature learning and crowded proposal filtering. WDPL extracts global and dense information at multilevel feature maps for each pathway rather than focusing on specific parts of the image. To further enhance the standard Region Proposal Network (RPN), Weighted Dense-Proposal Head (WDP Head) in WDPL is developed to measure the stability of each valid proposal. Thus, WDPL allows for a better representation of high-density objects in remote sensing images.

To tackle the second issue, flexible and novel modules are devised separately for the teacher and student networks, taking into account the differences in their training strategies and task requirements. These differences include variations in data augmentation and feature information. For the teacher network, we introduce the EM-based Fitting Policy (EFP) that differs from SoftTeacher [19], which splits images into supervised and unsupervised groups. EFP considers ground truth, supervised, and unsupervised images as inputs and maps them to the unsupervised prediction space without sophisticated modules. In each iteration, EFP employs the latent variables to automatically extract pseudo-label thresholds for consistency training.

Since geospatial instances exhibit large-scale variation, student predictions exacerbate the inconsistency associated with the second issue. For the student network, we introduce the Mean-Consistency-based Proposal Pruning (MCPP) to learn the consistency with EFP.

In general, we focus on partially labeled data rather than fully labeled data and propose a semi-supervised object detection method called Prospective Consistency Teacher (PCT). Unlike generic SSOD methods, PCT effectively mitigates the issues of inconsistencies arising from remote sensing targets. The concept map of the PCT unsupervised branch is illustrated in Fig. 2 (c). Our three major contributions are summarized as follows:

1) WDPL is proposed to enhance high-quality predictions proportion, strengthening the sensitivity of instances with multi-scale variations and significantly contributing to precise and robust bounding boxes.
2) Based on latent variables, EFP is designed to remove ambiguous and invalid pseudo boxes. The consistency of high-quality pseudo boxes is noticeably increased.
3) With the inconsistency of the teacher network reduced, we introduce a simple and effective approach named MCPP to select consistent proposals from the student network.

By training a limited amount of labeled data, our approach performs better on two public datasets: DOTA and DIOR [27]. In extensive experiments, it demonstrates structural robustness that PCT can be employed in different models type.

The rest of this paper is structured as follows. Section II introduces the related work from remote sensing object detection and teacher-student framework. Section The proposed method is described in Section III. Section IV reports the comprehensive experiments and analysis in detail. The last Section V summarizes the conclusion of our work.

II. RELATED WORK

With the development of Earth Vision technology, data-driven remote sensing object detection has been extensively investigated. SSOD is employed to mitigate models' reliance on a considerable quantity of labeled data. This section provides a succinct overview of object detection in remote sensing images. The subsequent discussion revolves around the teacher-student network for standard SSOD methods, concentrating on pseudo-based and consistency-based strategies.

A. Object Detection in Remote Sensing Images

Remote sensing images typically contain multi-scale and crowded objects, as opposed to natural image datasets like MS-COCO. CNN network outperforms most previous methods [28] by a significant margin. Object detectors are usually divided into two groups: single-stage-based and two-stage-based detectors. Then, we review the main contributions based on each of these two detectors.

Some methods explore single-stage-based networks for remote sensing images. Inspired by deformable ConvNets [29], Han et al. [2] introduce a S^2A-Net to align anchors and features. Yang et al. [30] design a smooth label mechanism named DCL for the angle classification, [31] rethinks the boundary problem of boxes to increase the fault tolerance of angles. However, this work still needs extensive data for training to obtain higher accuracy.

In contrast, two-stage-based detectors typically offer higher accuracy at the cost of slower processing speed. The DOTA dataset, introduced in [16], was the first to utilize a network based on oriented bounding boxes for object detection in remote sensing scenarios, which contain less background noise than images with horizontal bounding boxes. In [32], common non-maximum suppression (NMS) is replaced by Sig-NMS to improve detection accuracy for small targets in Very-High-Resolution (VHR) images. Ding et al. [33] reconstruct the RoI with transformer modules to exploit spatial transformations. CAD-Net [34] is proposed to employ the attention-based features by learning global and local contexts of various instances to study the object appearance differences. However, the attention-based methods provide the network with a broader view while sacrificing partial local perception, disrupting local and global contexts. Inspired by [35], our work first explores the features of multi-scale targets, then we sort and filter the dense proposals in region proposal networks.

B. Teacher-student Network in SSOD

Of late, the teacher-student framework [36], [37], [38], [39], [40], [41], [42], [43], [18] is widely adopted in semi-supervised object detection. The teacher-student network in SSOD primarily employs the Self-distillation paradigm, where the student and teacher networks share similar modules with different weights, such as Faster-RCNN [44]. The teacher network generates pseudo labels as ground truth for the student network. For most teacher-student networks, the weights of the student network are updated through backpropagation in the unsupervised branch. Exponential Moving Average [26] is used to update the weights of the teacher network with the student network after iteration. We will introduce the majority of work that focuses on two aspects: pseudo labels and consistency learning.

The pseudo-based SSOD methods aim at improving the quality of pseudo boxes. Li et al. [40] introduce certainty-aware pseudo labels which rethink the classification and localization quality respectively, yet it is complex and limited. MixMatch [45] works to generate low-entropy labels by MixUp process in data augmentation. STAC [18] designs a complex pseudo labels strategy by substantial unlabeled data augmentation.

The consistency-learning-based methods are mainly designed for data augmentation and loss function. By flipping each sample, Jeong et al. [46] compute the consistency loss in a simple way. Miyato et al. [39] rethink the local perturbation around input samples and propose a virtual adversarial loss. In [47], SSM is proposed to paste the proposal from an unlabeled image onto a labeled image for proposal consistency learning. Li et al. [19] propose a mechanism named Soft Teacher, which combines pseudo labels and consistency learning methods. Wang et al. [48] present a systematic approach to mitigate inconsistency in natural images. However, previous research primarily concentrates on acquiring consistency through RCNN

models. The high density of objects in remote sensing images causes the model to produce a substantial amount of inconsistent information in the RPN. The constraints of extracting consistency in RCNN are noteworthy and fail to address the issue effectively at heads. Inspired by SoftTeacher[19], we introduce a novel Prospective Consistency Teacher to delve into the pseudo labels and consistency of multi-scale objects for the remote sensing SSOD task.

III. PROPOSED METHOD

Generic SSOD methods emphasize the consistency of the RCNN module but overlook the prospective consistency available in the entire network. This is exacerbated by the large-scale variations and high object density in remote sensing scenarios.

Here, we propose a novel approach called Prospective Consistency Teacher (PCT), which fully utilizes consistency from the unsupervised branch and adaptively adjusts the threshold of pseudo-labels. It is worth noting that two-stage-based detectors can produce noisy and imprecise proposals at RPN stage. Therefore, we devise Weighted Dense-Proposal Learning (WDPL) to select high-quality and effective proposals, resulting in more precise prediction head candidate areas. Once we have more accurate candidates of various sizes, we use an EM-based Fitting Policy (EFP) to adaptively adjust pseudo-label thresholds for the teacher. Furthermore, we utilize Mean-Consistency-based Proposal Pruning (MCPP) to address the noisy and misaligned inconsistency in the student network by ranking the reliability of each proposal.

This section is divided into four parts. Firstly, we will introduce the architecture overview in subsection 3. A. Secondly, we design WDPL in subsection 3. B. Thirdly, EFP is introduced in subsection 3. C. Finally, we provide a detailed description of MCPP in subsection 3. D.

A. Overview

The supervised branch of the PCT adopts the same methodology as the general RSOD approach, and utilizes the unsupervised branch's student detector to learn ground truth from the input image. In the unsupervised branch, the overall architecture of PCT is illustrated in 3. Image I undergoes data augmentation processes divided into strong and weak augmentation sets. The former comprises K image pre-processing operations, with k_s being randomly selected for the student network. The latter involves k_w operations.

The student and teacher detectors in PCT are each composed of four distinct components: (1) WDPL resizes the enhanced image I_o to a smaller view I_r, then adopts ResNet [49] as the backbone to extract 5 layers of features from each view; (2) Feature Pyramid Network (FPN) [50] fuses redundant and multi-layer features into single-layer features with deep channels; (3) As a replacement for the vanilla RPN Head, the Weighted Dense-Proposal Head (WDP Head) in WDPL dynamically evaluates per proposal with a Gaussian score, improving the consistency of the first stage; (4) The RCNN module infers various targets' final classification and regression with the detection head. Corresponding RCNN modules are designed for student and teacher networks to improve consistency. In detail, EFP automatically sets thresholds for filtering pseudo labels from the teacher RCNN module. To mitigate noisy and misaligned inconsistencies in the student network, we use MCPP to rank instance-level consistency per remote sensing image to screen reliable and varied scale proposals.

Before calculating the loss in the unsupervised branch, each pair of predictions from the student and teacher networks corresponds to different location information. To address this issue, we use the geometry transform module to transfer the pseudo labels from the WDP Head or RCNN Head to the space of the student predictions.

B. Weighted Dense-Proposal Learning

Compared to natural images, remote sensing object detection presents a higher level of complexity and variability. It may hinder the accurate detection and classification of small objects. Common SSOD methods generally analyze the consistency of detection boxes in the RCNN module, which fails to reduce the inconsistency in densely generated proposals. WDPL aims to gather dense features and select effective candidate boxes for remote sensing object detection. It first employs dual pathways to capture features at multiple scales simultaneously, further using the WDP Head effectively weighs the features to select the most relevant ones.

WDPL facilitates the production of abundant proposals through a process of image resizing in two parallel pipelines, as shown in Fig. 4. The image size in its original state is denoted by I_o. The backbone extracts 5 feature maps f_1-f_5 (f_x represents the size of this layer feature is 2^{-x} of the image size). To enhance the capture of various targets, we generate small-view semantic information by resizing the input image to a smaller view I_r. Obviously, the feature extraction from I_r effectively avoids missed detection of actual objects. The model exhibits a characteristic of image size adaption, thereby there is a only need for a network of independent weights to be applied. Through these, we produce a large number of proposals without additional modules and model parameters to strengthen the capture capability.

However, how to distill the effective ones from numerous features and proposals? The vanilla RPN Head decomposes predictions into a classification score that relies on cross-entropy loss and a regression box that employs L1 loss. Notably, the RPN Head cannot accurately ascertain the confidence associated with each bounding box, though it can successfully achieve this objective in the classification task. To find practical proposals, we redesign the Weighted Dense-Proposal Head rather than the vanilla RPN Head. The green part in Fig. 4 shows the corresponding loss function.

Specifically, the 4 coordinates $\mathrm{d} = (d_x, d_y, d_w, d_h)$ of each box regression can be modeled as a Gaussian model [35]. Each coordinate represents the offset with respect to the anchor. Following it, we redefine the coordinate d_x as the form of $\mathrm{d_x} = \mathcal{N}(d_x \mid m,\ \sigma^2)$ (likewise for $d_y,\ d_w,\ d_h$). In other words, the single coordinate of each box is the form:

$$\mathrm{p} \in \mathcal{N}(p \mid m_p,\ \sigma_p^2) \quad \mathrm{p} \in prediction \tag{1}$$

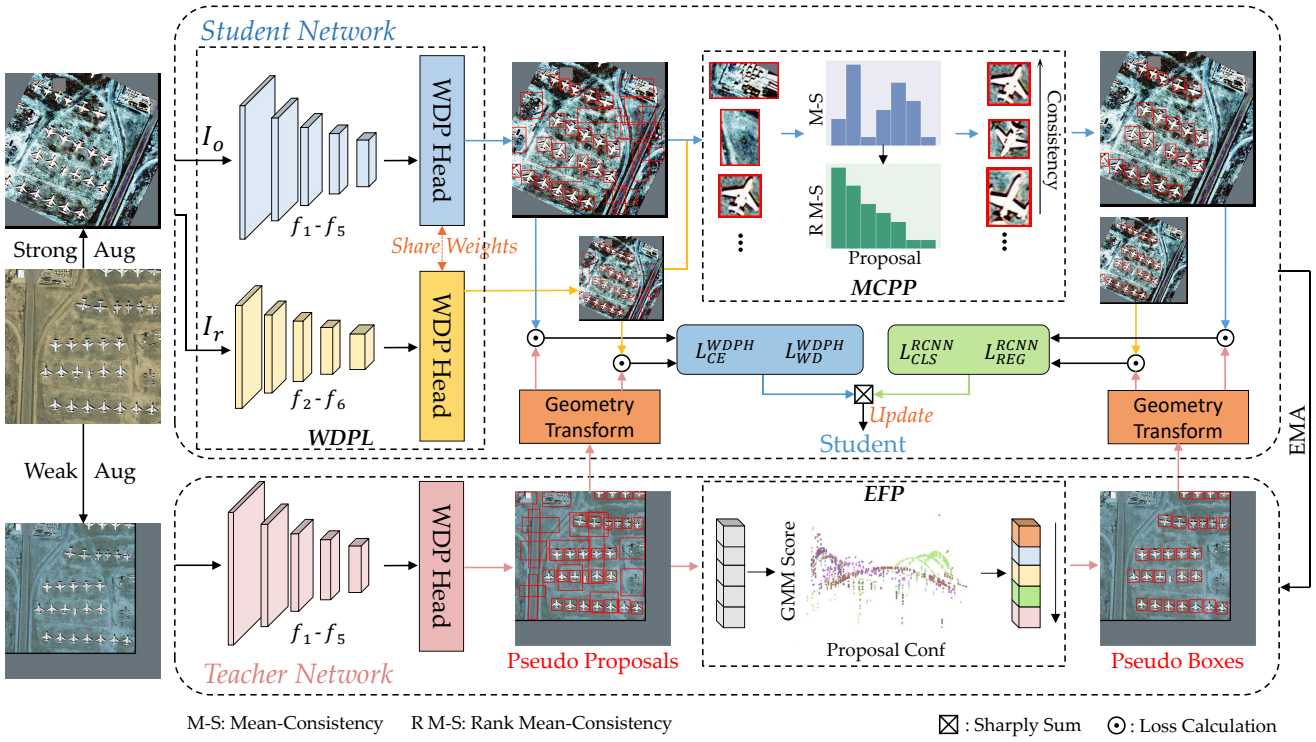

Fig. 3: The unsupervised branch of PCT. It contains student and teacher networks. The student network comprises WDPL and MCPP, while the EFP is designed for the teacher network.

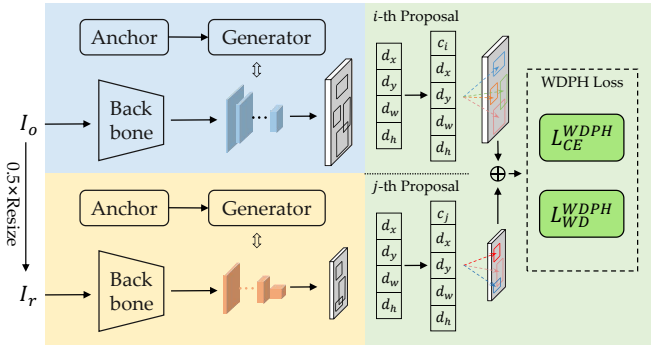

Fig. 4: The architecture of Weighted Dense-Proposal Learning. I_o is resized to half of its original size, denoted as I_r. A single student network is employed to process dual inputs and acquire multi-scale information. WDPL analyzes the confidence of each proposal and computes WDPL loss by merging these outcomes.

$$\mathrm{g} \in \mathcal{N}(g \mid m_g,\ \sigma_g^2), \quad \mathrm{g} \in ground\ truth \tag{2}$$

Here, p represents 4 coordinates in d of a predicted box. In [35], m_p consists of four values, and each value is the predicted coordinate of box. The variances σ_p^2 associated with each coordinate in d are transformed into a range between zero and one by applying a sigmoid function. Thus, σ_p^2 responds to the instability of the predicted box. While g is similar to p, it represents the ground truth box (*i.e.*, pseudo box). In this way, the expression for Gaussian cross-entropy can be written as:

$$GCE = \log \sigma_g + \frac{\sigma_p^2 + (m_p - m_g)^2}{2\sigma_g^2} + T \tag{3}$$

This equation demonstrates that each pseudo box's variance σ_g^2 is present in per sub-expression. Furthermore, σ_g^2 determines the magnitude of pseudo-label distribution. As σ_g^2 increases from 0 to $+\infty$, the distribution of pseudo-label transitions from Dirac Delta distribution to Uniform distribution. Therefore, the distributional factor β is designed to re-form variance. Focal loss [51] is proven to have high performance in dense object detection. We follow this form to GCE, and the Weighted dense-proposal loss is

$$L_{WD}^{WDPH} = FOCAL(\alpha) \cdot GCE(m_p, m_g, \sigma_p, \beta\sigma_g) \tag{4}$$

Here, L_{WD}^{WDPH} assigns confidence to each box as a weight, which alleviates performance fluctuations from dense objects. It should be noted that the ratio of α and β have a vital influence on the distribution of L_{WD}^{WDPH}. Combining the classification part, the unsupervised branch's loss is

$$L_{unsup}^{WDPH} = L_{CE}^{WDPH} + L_{WD}^{WDPH} \tag{5}$$

where L_{CE}^{WDPH} represents cross-entropy loss in the unsupervised branch for classification. L_{WD}^{WDPH} represents the weighted dense-proposal loss.

After estimating losses of RPN and RCNN separately, there will be an imbalance problem if we combine them directly for updating student network weights. Here, we introduce a sharp function that encourages predictions to be easily discriminated.

The ultimate loss of the student network is sharply sum and be defined as follows:

$$L_{unsup} = L_{unsup}^{WDPH} + \zeta \cdot L_{unsup}^{RCNN} \tag{6}$$

where L_{unsup}^{RCNN} is composed of L_{CLS}^{RCNN} and L_{REG}^{RCNN}, representing entire loss of RCNN in the unsupervised branch, ζ is the sharp factor.

C. EM-based Fitting Policy

Previous work mainly requires a fixed hyper-parameter to filter pseudo boxes with specific requirements. In this way, prediction confidences of the teacher model are different in datasets and model types. The static hyper-parameter has to be re-finetuned when the method meets another dataset or network schema (such as a different backbone). The EM-based Fitting Policy is devised to produce thresholds automatically to select valid pseudo boxes generated by the teacher model in the unsupervised branch.

Unlike previous work, we combine supervised samples (including labels) and unsupervised images as observable variables and ground truths for unsupervised images as invisible variables. Latent variables map observable variables directly to invisible variables without the need for other complicated operations, alleviating the computational complexity of models. The prediction confidence τ^{pb} is an expression of pseudo-box probabilities, so we fit it by typical maximum likelihood estimation: EM. Denote the observable and invisible variables as $x = (x_1, x_2, x_3, \cdots, x_n)$ and $z = (z_1, z_2, z_3, \cdots, z_n)$, respectively, and θ represents entire latent variables. We adopt these to the EM paradigm and obtain the likelihood function as the following equation:

$$LH(\tau_{j+1}^{pb}) = \sum_{i=1}^{n} \sum_{z_i} P(z_i|x_i, \theta_j) \log \frac{P(x_i, z_i; \theta)}{P(z_i|x_i, \theta_j)} \tag{7}$$

where $i \in [1, n]$ denotes $i\text{-}th$ dimension of observable or invisible variables, j denotes an iteration of EM algorithm, P denotes the actual distribution of variables.

Owing to the high robustness of the Gaussian distribution, we use the Gaussian mixture model as the probability distribution function. Naturally, the proposals produced by the teacher network, are divided into valid and invalid boxes. Valid boxes are essential for maintaining consistency in the model, and as such, they are retained. Conversely, invalid boxes must be automatically eliminated by the algorithm. Hence, we adopt two Gaussian distributions for the Gaussian mixture model. The function of the Gaussian mixture model is written as:

$$P(c) = \lambda_v \mathcal{N}(c_v|m_v, \sigma_v^2) + \lambda_{iv} \mathcal{N}(c_{iv}|m_{iv}, \sigma_{iv}^2) \tag{8}$$

where $\mathcal{N}(c_v|m_v, \sigma_v^2)$ and $\mathcal{N}(c_{iv}|m_{iv}, \sigma_{iv}^2)$ represent the confidence of valid and invalid boxes. λ_v and λ_{iv} denote probabilities belonging to the observable and invisible submodels, correspondingly.

D. Mean-Consistency-based Proposal Pruning

Within the unsupervised branch, pseudo boxes generated by the teacher detector are subject to filtration via an adaptive threshold approach, enhancing the models' performance by ensuring consistency in pseudo labels. Nevertheless, disregarding the selection of predictions made by the student network may give rise to inconsistent obstacles. Thus, it is worthy and necessary that impose constraints on the computation of RCNN loss during back-propagation.

MCPP is employed to refine proposals (*i.e.*, detected boxes) from the student network. The consistency of the detected boxes reflects the detection quality of the corresponding pseudo-boxes, while also aiding the student network in learning better consistency. Each pseudo box's consistency in the student network is formulated as:

$$\epsilon^i = {N_p}^{-1} \cdot \sum_{j=1}^{N_p} u_j^i \tag{9}$$

where u_j^i denotes the consistent value between the i-the valid pseudo box and the j-th detected box. ϵ^i denotes the $i\text{-}th$ pseudo box, N_p is a normalization factor, assigned to the number of positive proposals.

For simplicity, we denote u_j^i by IoU, which ranges from zero to one. The IoU value reflects the consistency between each pseudo box and the detected box. To assign the appropriate weight to each detected box, we measure the mean IoU of the different categories in each image and assign it to the corresponding detected boxes. Therefore, we use it as the instance-wise loss weight of detected boxes, which enables consistency learning in the student network.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

1) DOTA: It is a comprehensive and widely-used collection of remote sensing images intended for object detection. It comprises 2806 images ranging in size from 800*800 to 4k*4k, containing a total of 188282 instances. These instances belong to 15 object categories, namely 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 (HA), swimming pool (SP), and helicopter (HC).

2) DIOR: This dataset is another public and extensive collection, comprising 23463 images with 190288 instances. The images are of size 800*800. The dataset is divided into 20 categories, namely airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service areas, expressway toll station, golf field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill.

3) Metrics: The metrics mAP_{50} and mAP_{75} are widely used for evaluating the performance of remote sensing object detection methods across all categories. These metrics adopt Intersection over Union (IoU) thresholds of 0.5 and 0.75, respectively, to determine the accuracy of the detected objects.

In addition, $mAP_{50:95}$ is proposed by MS-COCO and uses 10 IoU thresholds from 0.5 to 0.95 linearly. It is another commonly used in the field of semi-supervised object detection.

TABLE I: ANALYSIS OF MULTI-SCALE DETECTION

Method	Dataset	Backbone	1%			5%			10%		
			mAP_s	mAP_m	mAP_l	mAP_s	mAP_m	mAP_l	mAP_s	mAP_m	mAP_l
supervised	DOTA	ResNet50	9.5	21.1	15.0	17.0	29.7	33.5	19.7	32.9	34.6
SoftTeacher			11.3	19.7	**17.4**	19.4	33.0	37.3	22.0	35.9	36.4
ours			**13.4**	**22.7**	17.1	**21.4**	**35.6**	**38.4**	**23.6**	**38.3**	**37.6**
supervised	DIOR	ResNet50	2.8	13.4	26.3	5.1	21.4	41.4	6.3	25.2	46.5
SoftTeacher			4.6	15.6	20.7	7.2	26.7	**45.6**	**9.2**	30.7	51.0
ours			**4.9**	**18.3**	**26.4**	**7.6**	**26.9**	45.4	8.5	**31.2**	**51.0**

TABLE II: PERFORMANCE OF SPECIFIC CATEGORY

Percentage	Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP_{50}
1%	supervised	61.4	25.2	15.1	12.5	33.9	54.3	66.5	72.9	0.0	**34.6**	9.5	22.0	41.3	23.9	12.0	32.3
	SoftTeacher	**77.0**	37.9	**25.9**	**21.0**	36.3	48.0	64.8	**90.8**	0.0	11.9	**29.3**	6.4	50.1	31.8	21.5	36.6
	ours	76.3	**39.2**	19.1	12.4	**40.6**	**62.4**	**69.5**	89.5	**1.3**	33.9	26.7	**29.4**	**52.0**	**36.1**	**25.3**	**40.9**
5%	supervised	77.7	49.9	28.9	33.4	48.2	70.6	72.1	86.4	40.2	62.2	37.0	45.9	64.7	37.3	30.9	52.4
	SoftTeacher	82.4	61.1	35.4	43.8	52.1	71.8	73.2	**91.7**	48.3	64.8	45.8	44.2	66.1	43.2	**33.3**	57.6
	ours	**82.9**	**63.3**	**36.3**	**46.3**	**54.3**	**78.4**	**74.0**	91.3	**54.9**	**71.8**	**53.3**	**46.3**	**67.6**	**48.9**	29.5	**60.2**
10%	supervised	80.9	54.8	36.3	40.4	53.3	76.8	73.3	90.9	44.1	62.1	48.8	49.9	69.4	43.5	3.0	55.2
	SoftTeacher	83.1	64.3	**45.9**	**51.6**	55.0	77.5	73.1	92.3	50.3	63.8	51.3	50.8	71.2	47.1	0.0	58.7
	ours	**84.4**	**66.6**	45.6	**51.6**	**55.3**	**81.1**	**74.5**	**92.8**	**58.7**	**71.2**	**53.9**	**51.7**	**72.0**	**50.9**	**14.0**	**61.7**

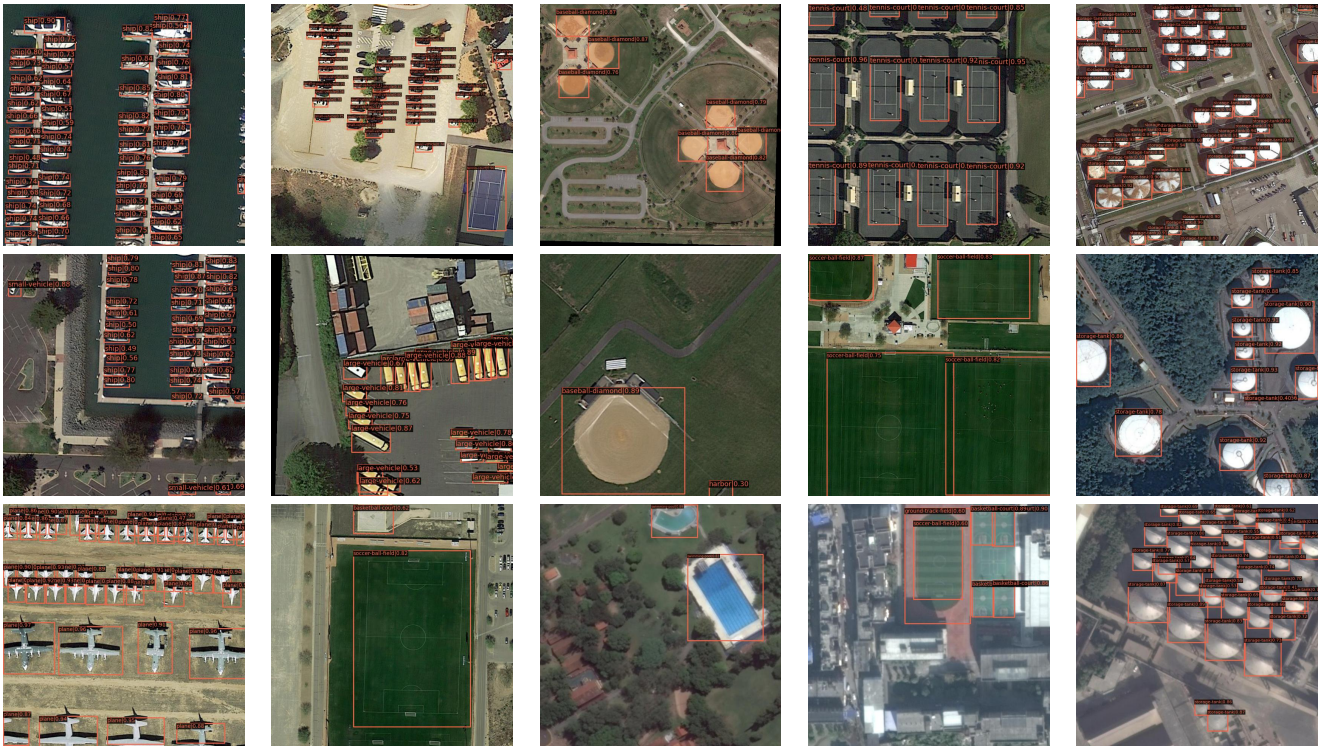

Fig. 5: Illustrations of student predictions on test images of DOTA. They concentrate on multi-scale and dense situations of remote sensing images.

Following MS-COCO format, all instances are divided into 3 categories based on the number of pixels, mAP_s evaluation metric corresponds to instances with areas smaller than 32^2. mAP_m pertains to instances with areas between 32^2 and 96^2, while mAP_l is concerned with instances larger than 96^2. In this way, our experiments comprehensively illustrate the evaluation of entire targets.

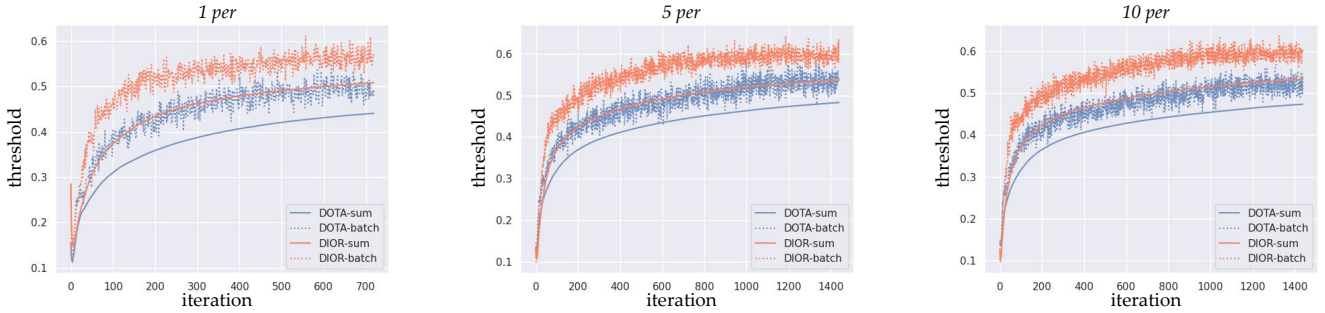

Fig. 6: Illustrations of thresholds during training. The "sum" mode represents the average of all thresholds from the start of training, the "batch" mode represents the mean of thresholds in each batch.

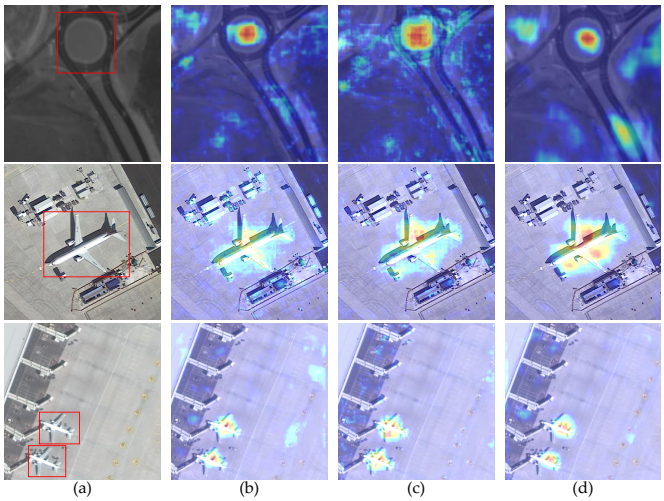

Fig. 7: Illustrations of heatmap at 10% protocol results. It displays the labeled image, PCT's backbone, PCT's neck, and SoftTeacher's backbone, represented as (a), (b), (c), and (d), respectively.

B. Implementation Details

Since deep semi-supervised learning of remote sensing object detection has yet to be widely studied, we mainly compare our method with the supervised model and select the horizontal bounding box as our task.

All experiments mentioned in this paper are implemented on NVIDIA RTX 3090 with a batch size of 5. To ensure a fair comparison, our method employs the same training and testing strategies as SoftTeacher, and we compared PCT with it in most of our experiments. For example, we utilize SGD as the optimizer with an initial learning rate of 0.0002, momentum of 0.9, and weight decay rate of 0.00001. Following the major semi-supervised object detection methods, we adopt Faster-RCNN as the student and teacher networks with separated weights. About augmentations, the k_w is set at 2, which contains a fixed set of random flip and resize operations. The image augmentation in K includes affine and color gamut transformations, with ten available options. We randomly select two transformations from these options and apply them to a single image. Additionally, we utilize erasing, cropping, and flipping techniques to enhance the image further.

TABLE III: COMPARISON WITH STATE-OF-THE-ART METHODS

Dataset	Method	1%		5%		10%	
		$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$	mAP_{50}
DOTA	supervised	17.6	32.3	29.5	52.4	31.6	55.2
	MeanTeacher [52]	20.2	39.4	30.1	53.3	31.5	53.8
	SoftTeacher [19]	19.0	36.6	33.1	57.6	34.1	58.7
	Consistent Teacher [48]	20.7	37.2	32.6	55.3	**36.5**	59.0
	ours	**20.8**	**40.9**	**34.5**	**60.2**	35.8	**61.7**
DIOR	supervised	16.4	29.5	26.9	48.5	31.2	54.9
	SoftTeacher	16.1	33.5	31.8	57.9	36.5	62.7
	ours	**20.2**	**40.9**	**32.4**	**58.6**	**36.4**	**63.3**

Following these, we set the value of k_s to 5.

In our experiments, we split the train set of DOTA and DIOR datasets into 1%, 5%, and 10% subsets according to the number of images with ground truths. The rest of each train set only takes image data without annotations. In this case, there are a few images with annotations and a vast amount of unlabeled images, which thoroughly tests the network's ability to use unlabeled data. The validation set of each dataset is employed to evaluate the performance of each method. Following a comprehensive experimental analysis in [19], the training strategy is executed for 180,000 iterations at 1% protocol and 360,000 iterations at 5% and 10% protocols.

C. Compare With the State-of-the-art Methods

This section conducts a comparative analysis of two remote sensing object detection datasets: DOTA and DIOR. SoftTeacher [19] and Consistent Teacher [48] are the recent state-of-the-art methods in generic SSOD. Since we employ the same data augmentation and other training settings as SoftTeacher, it is the primary point of comparison in all our experiments. $mAP_{50:95}$ is a fundamental metric for SSOD, while mAP_{50} is widely used in RSOD. These comprehensive detection metrics are employed to evaluate all methods in this section.

1) Comparison on DOTA: Table III suggests that PCT surpasses other methods on the DOTA dataset. It has demonstrated a stronger competitive performance for remote sensing images and enhances nearly all sub-protocol experiments. Specifically, our proposed method achieves 20.8, 34.5, and 35.8 $mAP_{50:95}$ on the validation set with 1%, 5%, and 10% labeled images, surpassing the supervised method by +3.2, +5.0, and +4.2 $mAP_{50:95}$, respectively. The experimental results conclusively demonstrate the superior performance of

PCT over fully supervised methods and establish state-of-the-art performance. Furthermore, under the metric mAP_{50}, our method achieves superior performance to SoftTeacher, with improvements of 4.3, 2.6, and 3.0, respectively. The reason is that the trade-off of PCT leverages consistency from both the RPN and RCNN modules to enhance target perception in remote sensing scenes. In contrast, previous methods focus solely on consistency within RCNN modules.

At 10% protocol, Consistent Teacher exhibits high accuracy only for sparse targets and thus performs slightly better than PCT in terms of $mAP_{50:95}$. However, it cannot be adapted to dense scenes, resulting in poor achievement under mAP_{50}.

We hypothesize that WDPL performs a coarse screening of potential consistency from RPN, while MCPP and EFP refine candidate consistency within RCNN.

2) Comparison on DIOR: To assess the efficacy of our approach on diverse datasets, we conducted a comparative evaluation of DIOR. Table III demonstrates that the proposed method dramatically improves DIOR without fine-tuning. It is worth noting that PCT delivers up to ten percentage points in mAP_{50}, which performs better than DOTA. Under $mAP_{50:95}$ and mAP_{50}, PCT reveals enormous precision and potential, and it has an increase of more than 3% in trials with different ratios. In numerical terms, PCT at 10% protocol of DIOR has three percentage points AP_{50} improvement than in DOTA, the $mAP_{50:95}$ and mAP_{50} metrics are competitive for 1% and 5% protocols. Table III demonstrates the effectiveness of the prospective consistency learning scheme across various data distributions.

3) Analysis With Multi-scale Objects: We conducted an additional experiment on DOTA to evaluate the performance of PCT in detecting targets with multi-scale variations. The results in Table I show that PCT enables efficient learning of unlabeled data even with small annotated training data. Specifically, on 1% protocol, PCT achieves 13.4 mAP_s and 22.7 mAP_m, which is an improvement over SoftTeacher, but still 0.3 percentage points behind mAP_l. The limited number of labeled images may hinder the learning of the appearance features of large targets. On 10% protocol, our approach pays more attention to large objects while maintaining high performance for other sizes. Also, we evaluated the multi-scale object detection performance of DIOR and presented the results in Table I. The results indicate that PCT can produce improvements of 2-5% across all protocols. Notably, PCT performs equally or even better than DOTA, showcasing remarkable detection capabilities for larger targets. However, it should be noted that PCT's ability to detect smaller and medium-sized targets is relatively limited. This can be attributed to the long-tail problem present in DOTA dataset. Despite this, PCT depicts a stronger efficiency of detection for small and medium-sized objects than the supervised model, by a margin of more than 2 points. This is due to our approach boosting the capacity for recognizing objects of various scales.

4) Analysis With Specific Categories: Table II presents the additional experiments conducted for each particular category. PCT assesses the consistency of all categories in each module, achieving performance improvements of varying degrees across most categories. When only 1% of the labeled data is used, PCT may not be robust enough and may not fit well for categories with limited data. However, with a slight increase in labeled images, PCT demonstrates high levels of robustness and validity in remote sensing images.

TABLE IV: ABLATION STUDIES OF EACH COMPONENT IN PCT

WDFL	EFP	MCPP	1%		5%		10%	
			$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$	mAP_{50}
✘	✘	✘	16.7	32.2	26.5	45.3	29.9	50.4
✘	✔	✘	17.0	32.7	27.5	46.6	30.5	50.9
✘	✘	✔	16.9	32.5	27.3	46.9	30.4	50.7
✘	✔	✔	17.2	34.5	29.3	50.4	31.0	52.9
✔	✘	✘	19.0	37.8	32.4	58.1	33.4	59.0
✔	✘	✔	19.1	38.6	32.6	57.9	34.8	60.3
✔	✔	✘	19.6	38.5	33.1	58.9	34.5	60.5
✔	✔	✔	**20.8**	**40.9**	**34.5**	**60.2**	**35.8**	**61.7**

D. Ablation Studies

Here, we conduct sufficient studies on DOTA dataset to evaluate the effectiveness and robustness of PCT. We adopt the same training strategies and test settings as the previous comparison experiments for standard SSOD methods to ensure a fair comparison. Specifically, we randomly divide the DOTA training set into three subsets, containing 1%, 5%, and 10% of the labeled images, respectively, while treating the remaining training set as unlabeled data.

1) Analysis of Each Module: The initial row of Table IV stands for the original teacher-student network, excluding WDPL, EFP, and MCPP. It performs worse than the supervised model, which uses the Faster-RCNN as the single detector. We hypothesize that inconsistent information exists in both the student and teacher networks, which affects efficacy when using only EFP or MCPP.

When the student and teacher networks adopt MCPP and EFP individually, there is less inconsistency in features and predictions, thus the mAP_{50} of three sub-experiments have more than 3 points improvements. Nevertheless, dense objects make the vanilla RPN generate a vast number of non-selected proposals, which is the fundamental source of the major inconsistency.

The integration of WDPL into the original teacher-student network leads to a noteworthy enhancement in the ultimate performance, as demonstrated by the observed increases of 2.3, 5.9, and 3.5 points at $mAP_{50:95}$ metric. It implies that the potential inconsistency in RPN has been properly eliminated, which is severely neglected by generic SSOD methods.

The penultimate and third lines demonstrate the effect of applying EFP and MCPP to WDPL respectively. It is notable that adding just one module did improve performance directly, but the effect was not impressive. However, EFP allows the PCT to adaptively adjust hyper parameters across multiple datasets to maintain robust capabilities.

Finally, the algorithm achieves the best performance when we add all the modules, which improves 8.7, 14.9 and 11.3 at mAP_{50}.

2) Analysis of WDPL Hyper Parameters: In Weighted Dense-Proposal Learning (WDPL), the ratio of hyper-parameters α and β is critical for learning from both labeled

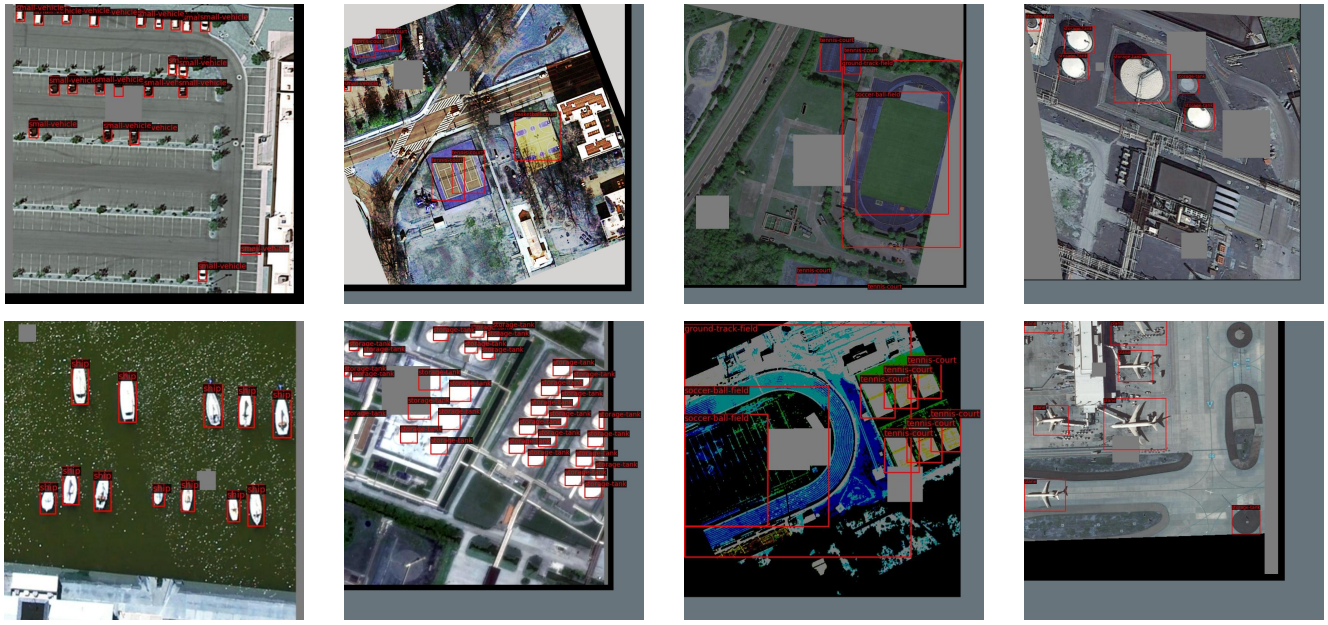

Fig. 8: Illustrations of student RCNN predictions via strong augmentation with 10% labeled train images.

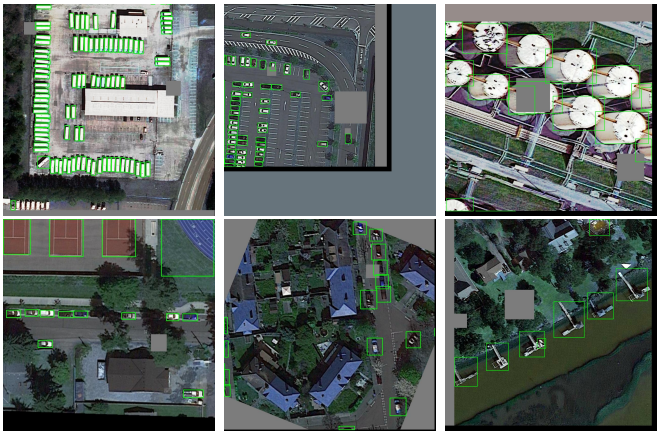

Fig. 9: Illustrations of student WDPL proposals via strong augmentation with 10% labeled train images. It focuses on the effectiveness of WDPL in dense scenes.

and unlabeled data. The ratio of the two values represents the degree of uncertainty associated with the pseudo label. We observed that slight changes in the ratio of α/β led to significant fluctuations in the final performance evaluation of the method. To determine the optimal solution, we selected a slicing unit of 0.5, resulting in a set of ratio values of [1.0, 1.5, 2.0, 2.5, 3.0]. The results presented in Table V indicate that the proposed method achieves peak performance when the ratio of α and β is set to 1.0. Notably, on 5% protocol, our method improves the $mAP_{50:95}$ by 3.6 (4.4 mAP_{50}) and increases the mAP by at least one point at other protocols.

3) Analysis of EFP Hyper Parameter: The high density of real targets may result in the generation of hundreds of proposals for a single image, potentially hampering the model's ability to detect the target and reducing the inference speed

TABLE V: ANALYSIS OF WDFL HYPER PARAMETERS ON DOTA

ratio	1%		5%		10%	
	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$	mAP_{50}
1.0	**20.8**	**40.9**	**34.5**	**60.2**	**35.8**	**61.7**
1.5	20.1	39.7	30.9	55.8	33.8	59.0
2.0	20.0	39.2	32.9	58.5	34.0	59.5
2.5	19.9	39.7	31.0	56.8	33.2	58.7
3.0	20.2	39.7	32.5	58.3	33.5	59.0

of our method. We employ a simple yet robust approach by sorting the proposals for each category based on their classification score and selecting only the top $scores_{num}$ proposals to input into the EFP. Table VI presents results for 10% protocol, revealing the model's overall solid performance and robustness when the EFP hyper parameter is set to a more significant value, as evidenced by minor performance fluctuations. Additionally, other experimental results have demonstrated the adaptive learning capability of EFP in various models and network architectures.

TABLE VI: ANALYSIS OF EFP HYPER PARAMETER ON DOTA

$scores_{num}$	10% percent set					
	$mAP_{50:95}$	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
25	32.3	57.3	32.0	21.1	34.1	33.8
50	32.7	58.2	32.3	22.6	34.9	35.4
75	**35.8**	**61.7**	**36.2**	**23.6**	**38.3**	**37.6**
100	34.8	60.4	35.1	23.2	37.2	36.8
125	34.7	60.1	35.3	22.2	36.6	36.5

4) Analysis of Swin Backbone: Previous experiments have utilized a single type of backbone, which may not adequately establish the effectiveness of diverse model architectures. We present an extension incorporating various backbones to gain

TABLE VII: COMPARISON ON SWIN TRANSFORMER BACKBONE

Backbone type	1%		5%		10%	
	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$	mAP_{50}
resnet50(sup)	17.6	32.3	29.5	52.4	31.6	55.2
resnet50	**20.8**	**40.9**	**34.5**	**60.2**	**35.8**	**61.7**
swin(sup)	18.8	36.3	30.4	56.1	31.5	56.5
swin	**21.2**	**41.3**	**31.9**	**59.1**	**35.7**	**62.3**

a more comprehensive understanding of PCT. ResNet is a commonly used CNN model for object detection tasks, while Transformer is an alternative framework employed in image classification and object detection tasks. Swin Transformer, in particular, has achieved state-of-the-art results in these fields [53]. In order to evaluate the robustness of our approach, we substitute ResNet with Swin Transformer while keeping all other components unchanged. To maintain fairness in this comparison, we adopt the same settings as those utilized in the ablation studies.

5) Visualization of Predictions in Test Set: Some detection results on test images are presented in Fig. 5. We trained PCT with only 10% labeled images from the DOTA train set, and the illustrations mainly focus on dense and multi-scale objects in the validation set of DOTA for the test task. Dense objects in DOTA include planes, small vehicles, and ships, which exhibit multi-scale characteristics, as do other categories, such as baseball diamonds and storage tanks. Overall, the results shown in Fig. 5 demonstrate that our method achieves impressive performance with only 10% labeled data.

6) Visualization of EFP Adaptive Thresholds: To verify the validity of EFP across various protocols and dataset distributions, we visualized the threshold movement during training. As shown in Fig. 6, the batch threshold represents the mean value of a batch of thresholds, which exhibits a violent oscillation. Hence, we compute the mean of all image thresholds from the start of training, which generally indicates the growing trend of the thresholds. These illustrations demonstrate the significance of EFP across datasets and protocols. Noteworthy is that thresholds on DOTA are generally about 0.1 lower than those in DIOR among various protocols. As a result of the increase in labeled data, upstream modules of PCT are improved to generate more accurate predictions. Compared with the 1% protocol, thresholds at 5% and 10% protocols are slightly higher among various datasets.

7) Visualization of Feature Heatmaps: To monitor the model's ability to capture the target, we visualized the heatmap of the backbone and neck in the student network, as shown in Fig. 7. SoftTeacher may be susceptible to interference from the background, whereas our method is designed to focus better on the target. Our method clearly demonstrates an excellent representational capacity for targets of varying scales.

8) Visualization of predictions in train set: Apart from feature heatmaps, we provide partial detection results to illustrate the efficiency of WDP and RCNN Head. Fig. 9 illustrates some dense proposal results of WDP Head, which only distinguishes between foreground and background instead of exact categories. Fig. 9 demonstrates the efficacy of WDPL by improving the quality of predictions from the RPN in dense-target scenes. Moreover, Fig. 8 illustrates ultimate box predictions of the student network. The above illustrations are computed by the student with 10% labeled and 90% unlabeled images.

V. CONCLUSION

This paper presents a novel semi-supervised object detection approach, PCT, using a teacher-student network for remote sensing imagery, which only utilizes 1%, 5%, and 10% of labeled data during training. The proposed PCT network consists of three main components, namely Weighted Dense-Proposal Learning (WDPL), EM-based Fitting Policy (EFP), and Mean-Consistency-based Proposal Pruning (MCPP). WDPL is responsible for enhancing the quality of dense proposals by assigning confidence to each box, thereby improving the detection accuracy in the Region Proposal Network (RPN) stage. Furthermore, EFP is designed to select pseudo boxes automatically, while MCPP filters dense proposals from the student network to facilitate consistency learning. Unlike prior work in natural scenes, *i.e.*, generic SSOD, our approach boosts consistency regularization across regions to better match dense and multi-scale geospatial objects. We have conducted numerous experiments to demonstrate the effectiveness and robustness of our method on the DOTA and DIOR datasets. PCT has achieved a mAP_{50} of 40.9, 60.2, and 61.7 with only 1%, 5%, and 10% of the labeled trainset on DOTA, and 40.9, 58.6, and 63.3 mAP_{50} on DIOR.

REFERENCES

[1] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "Abnet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[2] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[3] C. Zhang, K.-M. Lam, and Q. Wang, "Cof-net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[4] G. Wang, Y. Zhuang, H. Chen, X. Liu, T. Zhang, L. Li, S. Dong, and Q. Sang, "Fsod-net: Full-scale object detection from optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[5] Z. Zhao and X. Li, "Deformable density estimation via adaptive representation," *IEEE Transactions on Image Processing*, vol. 32, pp. 1134–1144, 2023.

[6] Z. Xiong, Y. Yuan, and Q. Wang, "Ask: Adaptively selecting key local features for rgb-d scene recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2722–2733, 2021.

[7] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7778–7796, 2022.

[8] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[9] Y. Yuan, X. He, and Z. Jiang, "Adaptive open domain recognition by coarse-to-fine prototype-based network," *Pattern Recognition*, vol. 128, p. 108657, 2022.

[10] Y. Yao, G. Cheng, G. Wang, S. Li, P. Zhou, X. Xie, and J. Han, "On improving bounding box representations for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[11] Y. Yuan and Y. Zhang, "Olcn: An optimized low coupling network for small objects detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[12] M. Zhang, Q. Li, Y. Miao, Y. Yuan, and Q. Wang, "Difference-guided aggregation network with multi-image pixel contrast for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[13] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2168–2187, 2022.

[14] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2022.

[15] L.-Z. Guo and Y.-F. Li, "Class-imbalanced semi-supervised learning with adaptive thresholding," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8082–8094.

[16] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[18] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020.

[19] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3060–3069.

[20] F. Zhang, T. Pan, and B. Wang, "Semi-supervised object detection with adaptive class-rebalancing self-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3252–3261.

[21] H. Li, Z. Wu, A. Shrivastava, and L. S. Davis, "Rethinking pseudo labels for semi-supervised object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1314–1322.

[22] C. Qin, L. Wang, Q. Ma, Y. Yin, H. Wang, and Y. Fu, "Semi-supervised domain adaptive structure learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 7179–7190, 2022.

[23] R. Yasarla, V. A. Sindagi, and V. M. Patel, "Semi-supervised image deraining using gaussian processes," *IEEE Transactions on Image Processing*, vol. 30, pp. 6570–6582, 2021.

[24] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3048–3068, 2022.

[25] H. Zhou, Z. Ge, S. Liu, W. Mao, Z. Li, H. Yu, and J. Sun, "Dense teacher: Dense pseudo-labels for semi-supervised object detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 2022, pp. 35–50.

[26] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[27] 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 journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

[28] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[29] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[30] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 819–15 829.

[31] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*. Springer, 2020, pp. 677–694.

[32] 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 Transactions on Geoscience and Remote Sensing*, 2019.

[33] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2849–2858.

[34] G. Zhang, S. Lu, and W. Zhang, "Cad-net: A context-aware detection network for objects in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.
[35] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 502–511.
[36] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *Advances in neural information processing systems*, vol. 27, 2014.
[37] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4119–4128.
[38] W. Lin and A. B. Chan, "Optimal transport minimization: Crowd localization on density maps for semi-supervised counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 663–21 673.
[39] T. Miyato, S. ichi Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[40] H. Li, Z. Wu, A. Shrivastava, and L. S. Davis, "Rethinking pseudo labels for semi-supervised object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1314–1322.
[41] Y. Li, D. Huang, D. Qin, L. Wang, and B. Gong, "Improving object detection with selective self-supervised self-training," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX*. Springer, 2020, pp. 589–607.
[42] W. Lin, J. Gao, Q. Wang, and X. Li, "Learning to detect anomaly events in crowd scenes from synthetic data," *Neurocomputing*, vol. 436, pp. 248–259, 2021.
[43] J. Wang, J. Gao, Y. Yuan, and Q. Wang, "Crowd localization from gaussian mixture scoped knowledge and scoped teacher," *IEEE Transactions on Image Processing*, vol. 32, pp. 1802–1814, 2023.
[44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, p. 1137, 2017.
[45] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.
[46] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," *Advances in neural information processing systems*, vol. 32, 2019.
[47] K. Wang, X. Yan, D. Zhang, L. Zhang, and L. Lin, "Towards human-machine cooperation: Self-supervised sample mining for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1605–1613.
[48] X. Wang, X. Yang, S. Zhang, Y. Li, L. Feng, S. Fang, C. Lyu, K. Chen, and W. Zhang, "Consistent targets provide better supervision in semi-supervised object detection," *arXiv preprint arXiv:2209.01589*, 2022.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[50] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
[52] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.
[53] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

Jinhao Shen received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2022. He is currently pursuing the M.S. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing and deep learning.

Cong Zhang received the Master's degree from the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. He is currently pursuing a Ph.D. with the Department of Electronic and Information Engineering at The Hong Kong Polytechnic University in Hong Kong. His research interests include computer vision and machine learning.

Yuan Yuan (Senior Member, IEEE) is currently a Full Professor with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or coauthored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS and *Pattern Recognition*, and the conference papers in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), British Machine Vision Conference (BMVC), International Conference on Image Processing (ICIP), and International Conference on Acoustics, Speech and Signal Processing (ICASSP). Her research interests include visual information processing and image/video content analysis.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.