RemoteDPL: A Semi-Supervised Object Detector With Dense Pseudo-Labels for Remote Sensing

Yongjie Ma, Xinyuan Zhou, Shiyong Lan, *Member, IEEE*, Wenwu Wang, *Senior Member, IEEE*, Zicheng Sun, and Yixin Qiao, *Graduate Student Member, IEEE*

***Abstract*— Deep learning-based object detection (OD) has seen substantial advancements; however, its practical deployment is often constrained by the need for large-scale labeled datasets. This limitation becomes even more critical in remote sensing imagery, where objects are densely distributed and exhibit significant scale variations. To address these challenges, we introduce dense pseudo-labels based SSOD method (RemoteDPL), a novel semi-supervised OD (SSOD) framework that leverages dense pseudo-labels (DPLs) and multiscale learning. RemoteDPL offers three key contributions. First, a fusion module is designed to dynamically integrate spatial and channel features across scales, improving detection across varied object sizes. Second, an instance density prediction branch is introduced to support pseudo-label mining, enhancing detection performance in densely populated regions. Finally, we propose a two-stage pseudo-label filtering strategy that first selects "pending" class predictions and then refines them using a joint confidence score based on both classification and density information. Extensive experiments on the DOTA-v1.0 and NWPU datasets confirm the effectiveness of RemoteDPL, demonstrating its clear advantage over existing state-of-the-art (SOTA) SSOD methods. On the NWPU dataset, RemoteDPL outperforms the SOTA baseline by +3.44%, +1.10%, and +1.62% under the settings of data labeled with 30%, 40%, and 50%, respectively, highlighting its strong capability in low-label remote sensing scenarios.**

***Index Terms*— Multiscale feature fusion, multiscale learning, pseudo-label mining, remote sensing image, semi-supervised OD (SSOD).**

I. Introduction

OBJECT detection (OD) has emerged as a critical component in remote sensing applications, facilitating the identification and analysis of objects within vast spatial imagery [1]. However, the development of robust object detectors hinges on the availability of extensive manually annotated data, a process that is not only labor-intensive but also time-consuming. In response to this challenge, semi-supervised techniques have gained interest, offering a promising avenue to alleviate the burden of manual annotation. Semi-supervised techniques are initially applied to image classification [2] and later extended to the general OD field. This approach, which enables the model to generate pseudo-label boxes without manual annotation, has greatly advanced the field of OD. Early work in this field includes the pioneering co-training algorithm [3], which leverages both labeled and unlabeled data by training two classifiers on distinct data views, using their agreement to generate annotations for the unlabeled samples. More recently, generative models, such as variational autoencoders (VAEs) [4] and generative adversarial networks (GANs) [5], have been used for semi-supervised learning (SSL). For example, VAEs have been used in [6] to generate synthetic data points to enhance the learning process, while GANs have been used in [7] to improve the robustness of the classifiers by creating realistic data samples.

Despite these advancements, SSL techniques still face substantial challenges in real-world applications, especially in the field of remote sensing. Remote sensing tasks often require algorithms to achieve scale invariance, as objects captured in satellite or aerial images vary significantly in size, ranging from large buildings to small vehicles. Furthermore, the dense distribution of objects in various scenes, such as complex urban areas, leads to object overlapping and occlusions, further complicating the OD and classification process. These distinct challenges in remote sensing make it more difficult to apply semi-supervised techniques effectively. Due to its high resolution and broad coverage, remote sensing imagery typically captures a wide variety of objects within a single image, ranging from large-scale structures like buildings and infrastructure to small entities such as vehicles and individuals. This scale diversity, as shown in Fig. 1(a), arises from various factors, such as the height variation of the capture platform.

In particular, a building captured by a high-altitude satellite may appear significantly different in size compared with the one captured by a low-flying drone. This heterogeneity in object scales presents unique challenges for standard semi-supervised OD (SSOD) models, which are often trained on datasets with relatively uniform object sizes. Anchor-based methods, such as Faster R-CNN [8], which are widely used for natural images, face challenges in adapting to various scales and densities of objects commonly found in remote sensing imagery. Selecting suitable anchor box sizes in advance becomes highly challenging because of the wide variability in object scales, as illustrated by our quantitative analysis of the DOTA-v1.0 dataset in Fig. 2(a), where the large scatter in object areas highlights the significant variation in scale. Moreover, anchor boxes of fixed-aspect ratio and scale are

Received 4 January 2025; revised 9 June 2025; accepted 16 June 2025. Date of publication 19 June 2025; date of current version 27 June 2025. This work was supported by the National Natural Science Foundation of China under Project 62371324. *(Yongjie Ma and Xinyuan Zhou contributed equally to this work.) (Corresponding author: Shiyong Lan.)*

Yongjie Ma, Xinyuan Zhou, Shiyong Lan, Zicheng Sun, and Yixin Qiao are with the College of Computer Science, Sichuan University, Chengdu 610064, China (e-mail: lanshiyong@scu.edu.cn).

Wenwu Wang is with the Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH Guildford, U.K.

Data is available on-line at https://github.com/SYLan2019/RemoteDPL.

Digital Object Identifier 10.1109/TGRS.2025.3581206

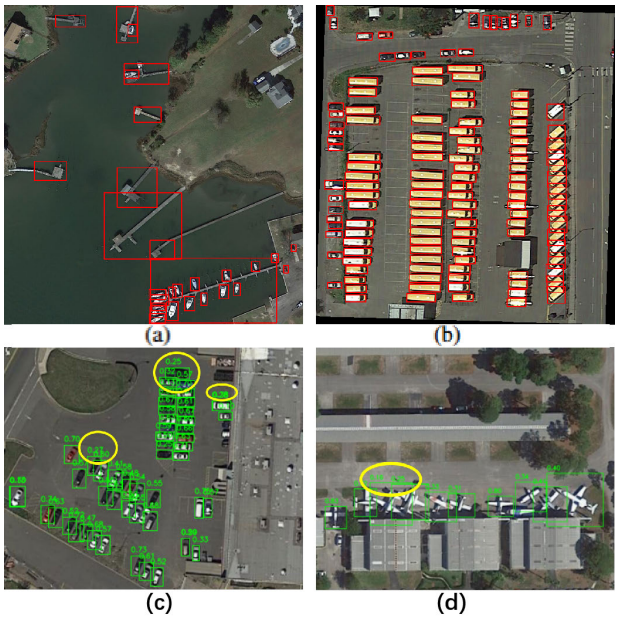

Fig. 1. (a) In remote sensing datasets, it is common for a single image to exhibit significant variations in instance scale. (b) Densely packed object distributions. Both challenges negatively affect detection performance; yet, current SSOD methods have not been specifically optimized to address them. (c) and (d) Limitations of single-threshold pseudo-label filtering. For instance, when the threshold is set to 0.3, valuable pseudo-labels within the yellow ellipsoids are discarded. This demonstrates the need for a more fine-grained pseudo-label filtering strategy to enhance the detection accuracy.

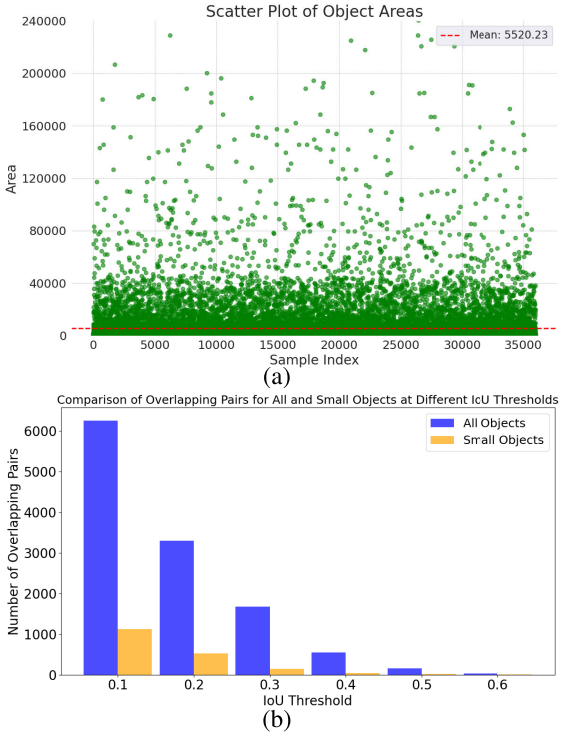

Fig. 2. Analysis of the DOTA-v1.0 dataset. (a) Scatter plot of object areas, highlighting the significant variation in object sizes, which pose challenges for scale-invariant detection. (b) Compares the number of overlapping object pairs at different IoU thresholds, revealing a significant reduction in overlap as the threshold increases, especially for small objects. These analyses underscore the complexity of object scale variations and spatial distribution within the dataset, thus demanding for robust detection models.

ill-suited for the varied and unpredictable object orientations in aerial and satellite imagery, often resulting in missed detections or imprecise bounding boxes. These unique characteristics necessitate the development of specialized SSOD methods capable of effectively handling extreme scale variations and ensuring robust performance in remote sensing applications.

Furthermore, dense instances present another formidable challenge in remote sensing applications. In scenarios with dense object distributions, as shown in Fig. 1(b), conventional postprocessing thresholds become less effective, leading to increased false positive rates [9]. One of the core challenges is the difficulty in distinguishing between closely spaced objects. Traditional nonmaximum suppression (NMS) algorithms [10], which are used to eliminate redundant bounding boxes, rely heavily on intersection over union (IoU) thresholds. When objects are densely packed, these IoU thresholds become less reliable. A high-IoU threshold can result in the retention of multiple boxes that overlap, leading to a spike in false positives. Conversely, with a low-IoU threshold, legitimate detections may be discarded, especially in cases where objects are partially occluded.

This challenge is further intensified by the inherent noise and the wide variation in object sizes within remote sensing imagery. As mentioned above, these images often feature objects of diverse scales and orientations. For example, in urban areas, buildings and vehicles may appear closely clustered, and elements like shadows or reflections can introduce false positives if not properly addressed. Consequently, traditional NMS techniques are often insufficient, highlighting the need for more advanced approaches to effectively manage dense and heterogeneous object distributions. As shown in Fig. 2(b), we analyze the overlapping boxes of DOTA-v1.0 at different IoU thresholds and find that the number of overlapping boxes decreases rapidly as the IoU threshold increases. This highlights how IoU threshold selection greatly impacts model performance in complex scenes, affecting the balance between accuracy and recall. It also underscores the added complexity in generating reliable pseudo-labels for SSOD.

To address these challenges, we introduce dense pseudo-labels (DPLs)-based SSOD method for remote sensing (RemoteDPL). Unlike previous semi-supervised approaches which often focused on natural scenes, dense pseudo-labels based SSOD method (RemoteDPL) aims at *the scale diversity and dense object distributions* that are typical in remote sensing imagery. RemoteDPL introduces a novel instance density estimation branch to examine feature maps across multiple scales and assess object density, thereby addressing the challenges posed by densely packed regions. Conventional thresholds often fail in such scenarios, leading to high-false positive rates [2], [10], whereas the density-driven approach of RemoteDPL selectively refines pseudo-labels in these challenging areas, ensuring more reliable supervision. Furthermore, we introduce an improvement in the integration of the hybrid-scale fusion branch, enhancing its ability to combine information from different scales effectively. By partitioning the feature maps into three distinct scales: original scale, downsampled scale, and hybrid scale, RemoteDPL goes beyond simple channel-level fusion [11], [12] by introducing a hybrid-scale fusion branch that explicitly integrates spatial

and channel information across different scales. This design enhances the ability of the model to handle large-scale variations typical in remote sensing imagery, leading to improved detection performance.

In addition, single-step thresholding is inadequate for complex scenes with high noise and dense, multiscale objects. To address this, RemoteDPL introduces a *two-stage filtering strategy* based on instance density estimation and hybrid-scale fusion. First, candidate pseudo-labels are filtered by confidence scores, separating high-confidence boxes from uncertain ones. The remaining labels are then refined using enhanced density and category scores from both hybrid- and original-scale features. This process reduces false positives and improves pseudo-label quality, boosting detection accuracy and robustness in dense remote sensing scenes. In summary, our main contributions are as follows.

1) We propose a two-stage filtering strategy that refines pseudo-labels using classification scores and object density cues. First, low-confidence boxes are adaptively filtered out. Then, the remaining candidates are evaluated using classification scores combined with object density estimates from hybrid-scale features.
2) An innovative density estimation branch is introduced to quantify the density score of objects within each region by analyzing hybrid-scale feature maps. Then, the density estimation branch is fused with the classification branch to improve the quality of pseudo-labels.
3) A novel fusion module is proposed to obtain the hybrid-scale feature representation, aggregating spatial and channel information from both the original and downsampled scale feature maps.
4) Extensive experiments conducted on DOTA-v1.0 and NWPU VHR-10 under various annotation ratios validate the superior performance of RemoteDPL compared with existing methods, and ablation studies confirm the effectiveness of its multiscale joint training, density estimation, and staged mining components.

This article serves as a comprehensive extension for its conference version [13]. Specifically, we provide a broader theoretical context in Sections I and II; we provide a more detailed and insightful analysis for density estimation, feature fusion, and two-stage filtering strategy; we thoroughly compare the experiments on more datasets to demonstrate the improved performance of the proposed model.

II. Related Works

A. Object Detection

OD is a foundational task in computer vision, critical for various applications. Existing detectors can be roughly categorized into anchor-based and anchor-free approaches. Anchor-based methods, such as YOLO [14], [15] and R-CNN series [8], [16], rely on predefined anchor boxes to guide detection. In contrast, anchor-free methods [17], [18], [19], [20], exemplified by FCOS [20], eliminate anchor-related hyperparameters, offering greater flexibility and robustness, particularly in scenarios with large-scale variations. Given the extreme scale diversity in remote sensing images, this article adopts the anchor-free FCOS [20] for SSL. By leveraging dense prediction grids and avoiding rigid anchor definitions, FCOS provides superior adaptability, making it well-suited for handling the significant scale variations of remote sensing imagery.

B. Semi-Supervised OD

SSL initially gained interest in the realm of image classification via leveraging unlabeled data to enhance the model performance [2]. Building upon their success in image classification, SSL-based OD methods have been proposed. For example, STAC [21] utilizes pseudo-labels generated from partially labeled data, combined with strong data augmentation. Unbiased Teacher [22] addresses the class imbalance issue via teacher–student mutual learning. ISMT [23] stabilizes pseudo-labels through NMS [24], whereas Instant-Teaching [25] uses Mixup and Mosaic for data augmentation. Humble Teacher [26] reduces noise with soft labels, while Dense Teacher (DenT) [27] refines localization through pixel-level pseudo-labels. DSL [28] introduces the adaptive filtering to promote scale invariance. MixTeacher [12] tackles scale variation with a mixed-scale teacher, and ARSL [29] employs IoU prediction to mitigate label selection ambiguity. VC [30], meanwhile, improves pseudo-label distribution alignment through virtual category mining. Despite these advances, most existing methods [25], [26], [27], [28], [29], [30] are devised for general image domains and overlook challenges specific to remote sensing, such as extreme multiscale variations and dense object distributions with frequent overlaps.

Recent studies have explored different structures to address scale variations. For example, the single-branch paradigm [as shown in Fig. 3(a)] relies on feature pyramids [21], [22], [23], [25], [26], [27], [29], [30], while the dual-branch paradigm [as shown in Fig. 3(b)] [28], [31], [32], [33] enforces consistency across multiple scales of the original input image. More recently, the triple-branch paradigm [as shown in Fig. 3(c)], first introduced in MixTeacher [12], combines features of the normal scale and downsampled scale to create a mixed scale representation. However, MixTeacher [12] emphasizes channel-level integration for natural images. In contrast, our RemoteDPL enhances the triple-branch framework by integrating spatial correlations and density estimation, tailored for remote sensing. This improves scalability and adaptability to extreme scale variation and dense object distributions, effectively addressing key limitations of prior semi-supervised methods.

C. SSOD in Remote Sensing

SSL has advanced remote sensing OD by leveraging both labeled and unlabeled data. Recent methods address key challenges, such as dense object layouts, large-scale variations, and high-annotation costs. For example, Chen et al. [34] utilized GANs [5] to generate additional training samples from unlabeled data, effectively improving detection accuracy through the data augmentation. However, the performance of such methods is highly dependent on the quality of the

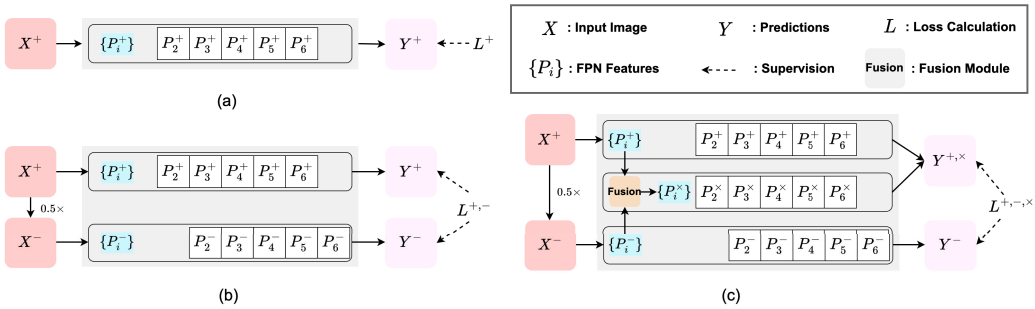

Fig. 3. Three architectural paradigms in SSOD are compared. (a) Single-branch paradigm, prevalent in mainstream semi-supervised detection models, primarily relies on feature pyramids to ensure scale invariance. (b) Dual-branch paradigm focuses on consistency predictions across images of different scales to achieve scale invariance. (c) Triple-branch paradigm employs a fusion module to integrate features from normal-scale and downsampled scales, creating mixed-scale features.

generated samples, which can vary significantly in complex remote sensing scenarios. Zhang et al. [35] combined active learning with SSL, using a teacher–student network and a region of interest comparison module (RoICM) to generate high-confidence pseudo-labels and select diverse samples for annotation, ensuring efficient use of labeled data. Despite its contributions, this approach emphasizes sample selection and pseudo-label generation but provides limited support for addressing the extreme scale variations inherent in remote sensing images. Wang et al. [36] proposed a weakly semi-supervised approach that integrates point-level annotations with bounding box labels, achieving competitive performance while reducing annotation costs. However, this method relies on precise point-level annotations as a prerequisite, which can limit its applicability in large-scale remote sensing datasets.

Other studies have focused on the detection of oriented (rotated) bounding boxes in remote sensing. For instance, Fu et al. [37] introduced a semi-supervised framework that combines multiview feature learning with rotational invariance constraints to improve the detection of arbitrarily oriented objects. Meanwhile, Fu et al. [38] utilized an adaptive teacher–student strategy to iteratively refine pseudo-labels with rotated box annotations, continuously enhancing detector robustness. Although effective for handling diverse orientations, these methods struggle with extreme scale variation and dense object distributions, and often require more complex architectures and annotation efforts.

While existing methods provide valuable insights, most address isolated challenges or use limited supervision, reducing their effectiveness in remote sensing. Given the extreme scale variation and dense object distribution, we propose a unified framework combining domain-specific adaptations with advanced SSL to better meet remote sensing needs.

III. Proposed Method

This section presents details of the proposed method, as illustrated in Fig. 4, with the FCOS [20] framework used as the baseline. FCOS, an anchor-free method, outperforms anchor-based approaches such as Faster R-CNN [8] in handling the scale variations common in remote sensing imagery, as it avoids reliance on predefined anchor sizes. Furthermore, we incorporate a pseudo-label-guided teacher–student framework to enhance detection and more effectively leverage unlabeled data. After each iteration, the teacher network is updated from the student network via exponential moving average (EMA) [39], ensuring more stable and reliable pseudo-label generation.

To address scale variations, we propose a multiscale fusion module that enhances scale invariance by integrating features across scales. Furthermore, we design a density estimation branch to predict instance density, which, combined with classification scores, forms a joint confidence score to assess pseudo-label reliability. A staged mining strategy further recovers missing pseudo-labels in dense regions, thus boosting detection performance. The detailed process is described in Algorithm 1. In Section III, we provide an in-depth discussion of the proposed RemoteDPL.

A. Multiscale Joint Training

Recent work [12] demonstrates that using downsampled views and enforcing consistency with regular inputs can enhance SSOD. For instance, MixTeacher [12] employs squeeze-and-excitation (SE) [11] [see Fig. 5(a)], improving multiscale detection. However, SE focuses on channelwise dependencies while neglecting spatial context, which may impact localization accuracy. To overcome this, we fuse both spatial and channel features from regular and downsampled views, improving robustness in multiscale training.

Drawing on the attention mechanisms of CBAM [40] and CA [41], we explore two distinct fusion strategies, as shown in Fig. 5(b) and (c), in a manner inspired by MixTeacher [12]. Different from CBAM and CA which refine the features through attention mechanisms, our proposed fusion module first combines the feature of the current layer P_i with that of the previous layer P_{i-1} (with different spatial resolutions), to obtain the features of multiple scales P_+. We then further enhance P_+ through the CBAM-like or CA-like module to obtain the multiscale feature P_{out}. Both fusion strategies are designed to enhance the mixed-scale feature pyramid and improve the model's ability to recognize objects at different scales. Specifically, the channel attention in the CBAM-like module adaptively adjusts the weights of concatenated multiscale features. A convolution then reduces these from $2C$ to C channels, further enhance the feature fusion. Similarly, the CA-like module performs multiscale fusion but incorporates

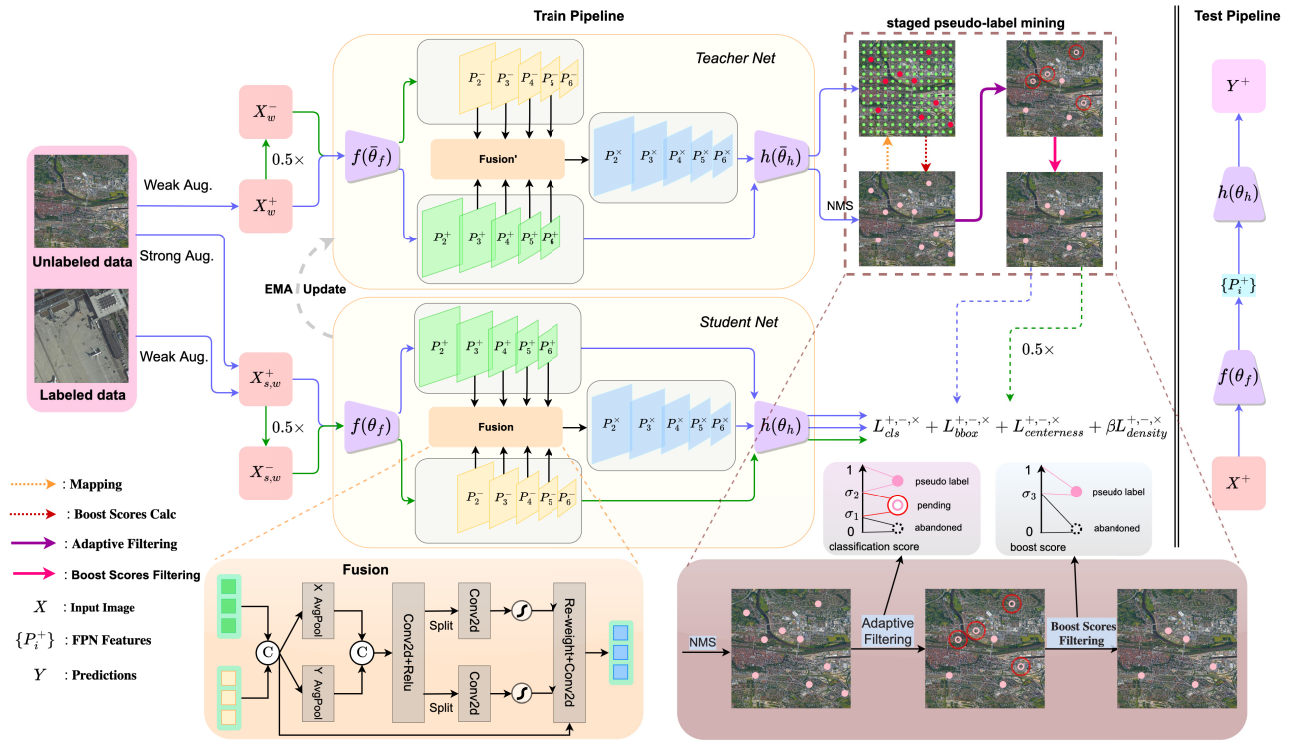

Fig. 4. Pipeline of our proposed method. During the training phase, we introduce an additional branch to predict instance density, forming a joint confidence score with the classification score for deep-level exploration of pseudo-labels. We employ multiscale learning by dividing the feature maps into the original scale, downsampled scale, and hybrid scale, which integrates spatial and channel information from the first two scales. After jointly training these three types of features, we update the teacher model using the student model with EMA.

positional awareness into channel attention, boosting spatial context aggregation. The CA-like fusion method [see Fig. 5(c)] is the default choice for fusion in our proposed model.

Our process involves downsampling the input image by a factor of 0 and feeding it into the network to acquire downsampled feature maps, denoted by $\mathcal{P}^- = \{P_2^-, \ldots, P_6^-\}$, capturing coarse-grained information. These features are then fused with the original-scale feature maps $\mathcal{P}^+ = \{P_2^+, \ldots, P_6^+\}$, using a module illustrated in Fig. 5(c), to obtain the hybrid-scale feature maps $\mathcal{P}^\times = \{P_2^\times, \ldots, P_6^\times\}$ as follows:

$$P_i^\times = f\left(P_i^+, P_{i-1}^-\right) \tag{1}$$

where the subscript i indexes the layer of the feature map output from the FPN network [42] and f represents the fusion operation, as illustrated in Fig. 5(c). This fusion process integrates channel and spatial feature maps of different scales, thereby enhancing the model's capability to recognize objects of different scales.

These hybrid-scale feature maps serve as the basis for the hybrid-scale branch during the training of the student model. Leveraging the rich information encapsulated within the hybrid-scale feature maps, we adopt a filtering approach during the pseudo-label generation phase. Specifically, we prioritize pseudo-labels based on the improvement in predictions made by the hybrid-scale branch compared with the original scale branch for the same predicted bounding box. This selective filtering ensures that only high-quality pseudo-labels are retained, thereby enhancing the overall robustness of the model.

In our training framework, we jointly train the original-scale, downsampled-scale, and hybrid-scale branches. This multiscale approach is designed to strengthen the model's ability to handle objects of varying sizes. The overall loss function governing the training process is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{unsup}} \tag{2}$$

where L_{sup} denotes the supervised loss computed on labeled images and L_{unsup} signifies the unsupervised loss calculated on unlabeled images. The parameter λ is used to balance the contributions of the supervised and unsupervised losses.

According to the usual settings in SSOD methods [25], [26], [27], [28], [29], [30], we also set λ to 2. Both the supervised loss L_{sup} and the unsupervised loss L_{unsup} are defined in (3). As in MixTeacher [12], we compute the total loss by summing the loss contributions from all scales

$$\begin{aligned}\mathcal{L}_t = \mathcal{L}_{t,\text{cls}}^{+,-,\times} + \mathcal{L}_{t,\text{bbox}}^{+,-,\times} + \mathcal{L}_{t,\text{centerness}}^{+,-,\times} \\ + \beta \mathcal{L}_{t,\text{density}}^{+,-,\times}, \quad t \in \{\text{sup,unsup}\}\end{aligned} \tag{3}$$

where L_{cls} and $L_{\text{centerness}}$ are the original FCOS focal loss and cross-entropy loss, respectively. The term L_{bbox} utilizes the complete IoU (CIoU) loss to measure the precision of the predictions of the boundary box. In addition, L_{density} employs the binary cross-entropy (BCE) loss for instance density estimation. The parameter β denotes the weighting factor applied to the density loss, balancing its contribution within the overall loss function.

In more detail, L_{cls} is responsible for the classification task, using the focal loss to address the class imbalance by focusing

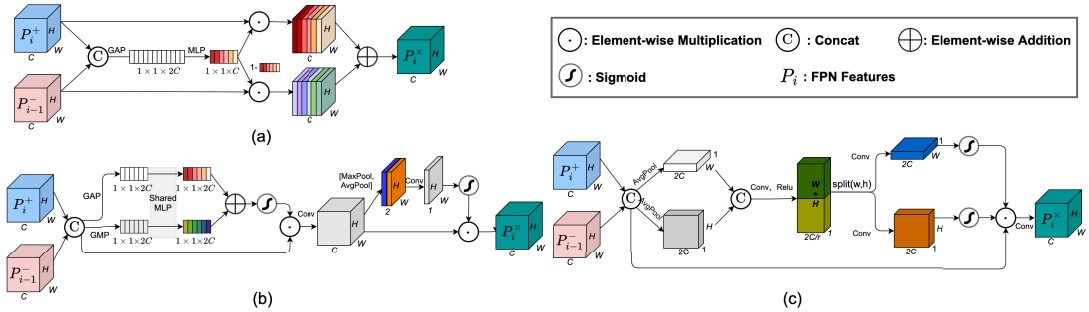

Fig. 5. (a) Fusion method in MixTeacher [12] serves as a baseline. (b) CBAM-like fusion module and (c) CA-like fusion module. (b) and (c) Adopted in our model can dynamically blend information from spatial and channel dimensions to accurately represent the hybrid-scale features, while baseline (a) used in MixTeacher only considers channel information.

Algorithm 1 Pseudocode of RemoteDPL

Framework: Teacher Model,Student Model,Fusion Module
Loss function: Classification Loss(L_{cls}),
Regression Loss(L_{bbox}),
Centerness Loss($L_{centerness}$),
Density Loss($L_{density}$)

Input:
Labeled Dataset $\mathcal{D}_l = \{(\mathcal{X}_l^i, \mathcal{Y}_l^i)|_{i=1}^{N_l}\}$
Unlabeled Dataset $\mathcal{D}_u = \{\mathcal{X}_u^i|_{i=1}^{N_u}\}$
Maximum Epoch K
Augmentation strategies (Weak Aug, Strong Aug)
Initial Pseudo-Label Generation Stage:
Initialize: Base detector with pre-trained weights
Train the base detector using labeled data $\mathcal{D}_l$
Use the trained detector to generate initial pseudo-labels for unlabeled data $\mathcal{Y}_{u0}$
Output 1: Unlabeled Dataset $\mathcal{D}_{u0} = \{\mathcal{X}_u, \mathcal{Y}_{u0}\}$
RemoteDPL Training Stage:
Initialize: Teacher Model(T) with pre-trained weights,
Student Model(S) with pre-trained weights
for all $i = 1, \ldots, \frac{K}{4}$ **do**
  for each batch of $(\mathcal{D}_{l0}, \mathcal{D}_{u0})$ **do**
    Weak augmentation is used for labeled data
    Strong augmentation for unlabeled data
    Update the student model S parameters according to Eq. (2)
    Update the teacher model T parameters using EMA: $\theta_t^j = \alpha\theta_t^{j-1} + (1-\alpha)\theta_s^j$
  end for
end for
for all $i = \frac{K}{4} + 1, \ldots, K$ **do**
  for each batch of $(\mathcal{D}_{li}, \mathcal{D}_{ui})$ **do**
    Weak augmentation is used for labeled data
    Strong augmentation for unlabeled data
    Update the student model S parameters according to Eq. (2)
    Update the teacher model T parameters using EMA: $\theta_t^j = \alpha\theta_t^{j-1} + (1-\alpha)\theta_s^j$
    Use the teacher model T to generate new pseudo-labels $\mathcal{Y}_{ui}$
  end for
end for
Output 2: Final trained Student Model (S)

more on hard-to-classify examples. $L_{\text{centerness}}$ enhances the model's ability to predict the likelihood that a pixel is within a range around the center of an object, thus refining the localization accuracy. The CIoU loss L_{bbox} not only considers the overlap between predicted and ground-truth bounding boxes but also takes into account the distance between their central points and the aspect ratio, providing a more holistic measure of bounding box quality. For L_{density}, the BCE loss is utilized to improve the prediction of instance density maps, which helps to identify dense regions within the image. The parameter β is crucial as it adjusts the impact of the density estimation task relative to other loss components, ensuring that the model maintains balance across all tasks.

By integrating these loss components, the overall loss function ensures that the model not only performs well in classification and localization but also effectively handles the complexity in object density and scale variations, thereby enhancing its robustness and accuracy in complex remote sensing scenarios.

B. Estimation of Object Distribution Density

In remote sensing, targets are often small and densely distributed, leading to a significant amount of overlapping instances, as illustrated in Fig. 1(b). Regions of high-instance density can increase the likelihood of detection errors, which adversely affect the quality of the generated pseudo-labels. Therefore, it is crucial for the model to dynamically adjust its optimization decisions during the postprocessing stage based on the instance density. This adaptability is essential to enhance the model's performance and applicability in complex and densely populated environments.

To address this challenge, we introduced a dedicated branch within the FCOS framework [20] to predict the instance density in the region of the sampled points, as shown in Fig. 6(a). Specifically, similar to the existing classification and regression branches in the FCOS head, we incorporated four additional convolutional layers following the FPN network [42] to form the density prediction branch. This branch is designed to estimate the density of instances within a given region, allowing the model to understand and adapt to varying densities. Using the detailed spatial and contextual information provided by the FPN layers [42], it enables the model to make more informed decisions based on the density of objects in the scene.

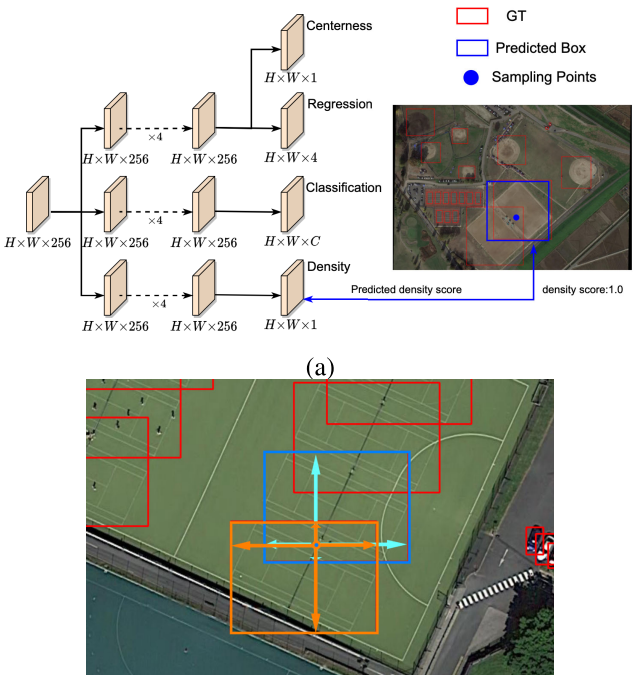

Fig. 6. (a) Density estimation branch is added to the original FCOS architecture to predict the density of sample points. (b) When a positive sample point is located between the two ground-truth boxes, the count of the ground-truth boxes represented by this positive sample point is 2. There are many positive sample points like this in an image. We take the maximum value of the ground-truth box count for these points as $n_{\max}$, and the density score for this point is $(2/n_{\max})$.

During the label assignment process, when determining whether the position (x, y) in the feature map corresponds to a positive sample point, we define the density as the count of this point within the corresponding ground-truth box, normalized at the image level, as depicted in Fig. 6(b). The density regression target is formulated as follows:

$$d^*_{x,y} = \frac{n_{x,y}}{\max\limits_{i,j} n_{i,j}} \tag{4}$$

where i and j denote the coordinates of the sample point on the feature map. The density value $d^*_{x,y}$, which is normalized to the range $[0, 1]$, represents the concentration of instances within the region corresponding to that sample point. A higher density value, i.e., closer to 1, signifies a greater abundance of instances in the vicinity. To train the density branch, we employ the BCE loss, which effectively captures the disparity between predicted and ground-truth density, guiding the model toward accurate density estimation.

During pseudo-label generation, our method incorporates density branch predictions to refine candidate boxes. For boxes with confidence scores below the positive threshold, we apply an additional density-based filter to identify high-quality pseudo-labels. This helps to prioritize boxes in dense regions, enhancing overall detection accuracy and pseudo-label reliability.

C. Two-Stage Filtering Strategy

In SSOD, improving the pseudo-label quality is crucial. To address this, we propose a strategy that leverages the first two modules for effective pseudo-label mining. During the pseudo-label generation, we compare confidence and density scores from hybrid- and original-scale branches to evaluate and refine potential pseudo-labels based on score improvements. By utilizing insights from both the hybrid-scale and original-scale branches, our strategy identifies pseudo-labels with significant improvements in confidence and density scores. This evaluation helps to ensure the selection of high-quality pseudo-labels, ultimately enhancing the accuracy and effectiveness of the SSOD framework.

1) Mapping and Boost Scores Calculation: During the pseudo-label generation phase, following postprocessing steps like NMS [24], we undertake a crucial step to enhance the quality of pseudo-labels. Specifically, for each feature point selected from the original-scale feature map, we match it to the same region on the hybrid-scale feature map. By mapping coordinates between the feature maps, we establish a direct comparison between predictions made at different scales. This mapping procedure enables us to calculate the score improvement of predicted boxes derived from the hybrid-scale branch relative to those generated by the original-scale branch.

To this end, we introduce a joint confidence score s to integrate the classification score p and the density score d as follows:

$$s = \sqrt{2\left(\alpha p^2 + (1-\alpha)d^2\right)} \tag{5}$$

where α is a hyperparameter, representing the combination ratio of p and d. Consequently, s is used to guide the selection of high-quality pseudo-labels, ensuring that only the most reliable predictions are retained for subsequent training stages. The final score improvement is formulated as follows:

$$\Delta s_{r_i} = \max\left(0, s^{\times}_{r_i} - s^{+}_{r_i}\right) \tag{6}$$

where r_i signifies the predicted boxes retained from the original-scale branch following postprocessing operations, $\times$ denotes the hybrid-scale branch, and $+$ indicates the original-scale branch. These improvement values serve as indicators for subsequent high-quality pseudo-label mining.

2) Adaptive Filtering and Boost Scores Filtering: We introduce a multistage filtering strategy, inspired by previous SSOD methods, such as dense learning (DSL) [28] and ambiguity-resistant SSL (ARSL) [29]. Initially, hierarchical filtering based on the confidence scores $p_{x,y}$ of pseudo-labels is conducted, where $p_{x,y}$ is the classification score for point (x, y) in the original-scale branch. We utilize two thresholds, σ_1 and σ_2, for the initial filtering of pseudo-labels

$$l_{x,y} = \begin{cases} \text{pseudo label}, & p_{x,y} \geq \sigma_2 \\ \text{pending}, & \sigma_1 \leq p_{x,y} < \sigma_2 \\ \text{abandoned}, & p_{x,y} < \sigma_1. \end{cases} \tag{7}$$

In our experimental setup, we referred to the category-adaptive filtering strategy of DSL [28] to set σ_1 and σ_2. Specifically, σ_1 is fixed at 0.1, while σ_2 follows an adaptive mechanism to be dynamically adjusted for each category based on the characteristics of a specific category.

The category-adaptive threshold is calculated as follows:

$$\sigma_2^k = \left(\frac{\sum_{x,y} \mathbb{1}_{\{l^*_{x,y}==k\}} l_{x,y}}{N_{\text{pos}}} \right)^{\eta} \tau \tag{8}$$

where $l^*_{x,y}$ represents the predicted category of the bounding box and k denotes the kth class in the total dataset categories C. Here, we set $\eta = 0.7$ and $\tau = 0.35$. Following the initial filtering stage, we proceed with a second filtering step based on the confidence boost values of the "pending" bounding boxes identified in the first filtering step. If the boost value Δs_{r_i} of the r_ith bounding box, as shown in (6), exceeds a threshold σ_3, we relabel/change this box as pseudo-label, where σ_3 is a hyperparameter.

Remarks: We begin by applying a classification confidence threshold σ_1 to filter out low-confidence boxes, effectively removing most noisy pseudo-labels. The remaining "pending" boxes, those with scores between σ_1 and a higher threshold σ_2, are further assessed using a "boost value" that integrates both density and classification scores. Although these boxes may have modest classification scores, a high-boost value indicates richer feature representation from the hybrid-scale branch, often revealing instances underestimated at the original scale. As a result, boxes with strong density signals or notable improvements from multiscale features are retained as valuable pseudo-labels.

IV. Experiments

A. Dataset and Metric

We conducted experiments on the DOTA-v1.0 [43] and NWPU VHR-10 [44] datasets, respectively.

1) DOTA-v1.0 Dataset: The DOTA-v1.0 dataset consists of 2806 images across 15 common categories, totaling 188 282 labeled instances. To create a comprehensive annotated dataset, we merge the training and validation sets of DOTA-v1.0. From this merged dataset, 80% is allocated for training and the remaining 20% for validation. The original test set of DOTA-v1.0 is then used as the unlabeled dataset. For labeled data, we sample from the combined training set at proportions of 1%, 5%, and 10% relative to the size of the unlabeled dataset. Due to the high resolution of DOTA images, we apply a cropping procedure. Each image is divided into patches of 1000 $\times$ 1000 pixels with a stride of 450 pixels. The annotations for the cropped patches are adjusted accordingly by removing entries with empty targets, targets outside image bounds, targets marked as difficult (difficulty value of 1), and targets that are too small to be reliably detected.

2) NWPU VHR-10 Dataset: NWPU VHR-10 [44] is a geospatial remote sensing dataset for aerial OD, featuring ten object categories. It consists of 650 annotated images, each containing at least one instance, with a total of 3651 target instances, and 150 background images without any targets. We divide the 650 annotated images into training (total annotated data), validation, and testing (unlabeled) sets, following a 3:2:5 ratio. This division aims to ensure that unlabeled data constitutes the majority, enabling us to sample labeled data from the total annotated set based on specific proportions relative to the unlabeled data. For our experiments, we extract the labeled data with proportions of 30%, 40%, and 50% of the unlabeled dataset.

In all experiments, we evaluated the performance of the model using mean average precision (mAP) on the overall validation set. This mAP metric, as defined in pascal visual object classes (Pascal VOC) challenge [45], has been used as a universal metric for evaluating OD accuracy.

B. Experimental Settings

Our model is built on the anchor-free FCOS detector [20], utilizing ResNet-50 as the backbone, FPN [42] as the neck, and dense heads for detection. During training, we apply weak augmentation to the labeled data and strong augmentation to the unlabeled data, while pseudo-label generation for unlabeled data uses weak augmentation. For weak augmentation, we employ random flipping, while strong augmentation, includes random flip, color jittering, cutout, and patch shuffle. The model is trained on two RTX 3090 GPUs for a total of 28 epochs. We use the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.0015, which is reduced by a factor of 10 at the 20th and 26th epochs. The momentum and weight decay are set to 0.9 and 0.0001, respectively.

We set the unsupervised loss rate λ to 2, following the general setting of existing SSOD frameworks [20], [25], [26], [27], [28], [29]. In addition, we fix the density score loss weight β at 0.1. For staged filtering, thresholds are set as follows: $\sigma_1 = 0.1$, $\sigma_3 = 0.65$, and σ_2 is defined in (8). On each GPU, we randomly sample two images from the labeled and unlabeled datasets in a 1:1 ratio. Consistent with prior SSOD studies [27], [28], we employ a "burn-in" strategy to initialize the teacher model.

C. Main Results

In this section, we compare our method with other state-of-the-art (SOTA) SSOD approaches in the general domain, including the methods from [22], [27], [28], and [29]. We perform evaluations for these methods using the DOTA-v1.0 and NWPU VHR-10 datasets. To ensure a fair comparison, we used a consistent number of training epochs for each method. In addition, we perform the same mAP evaluation across all methods, where mAP is computed by averaging the precision–recall curve area for each object class (with an IoU threshold of 0.5 for a detection to be considered correct). Furthermore, for fair comparison, we utilized mAP as a performance metric for all compared methods. Note that the SSOD algorithms in the remote sensing field as discussed in the related work Section II, to the best of the authors' knowledge, do not have publicly available open-source implementations. Furthermore, there is a lack of a unified standard for the dataset construction for SSOD in the remote sensing field. For these reasons, we only compared with the SOTA methods in the general domain, similar to most existing work for remote sensing detection.

1) DOTA-v1.0 Dataset: We first evaluate our method on the DOTA-v1.0 dataset under different ratios of labeled data,

TABLE I

EXPERIMENTAL RESULTS ON DOTA-V1.0 UNDER THE PARTIALLY/FULLY LABELED DATA SETTING. EXPERIMENTS ARE CONDUCTED ON 1%, 5%, 10%, AND 100% LABELED DATA. (·) DENOTES THE IMPROVEMENT COMPARED WITH THE BEST RESULT IN BASELINES

MODEL CATEGORY	METHODS	VENUE	1%	5%	10%	100%
SUPERVISED	FASTER R-CNN [8]	NEURIPS 2016	48.25	71.03	80.40	-
	FCOS [20]	ICCV 2019	47.93	69.31	79.57	-
SEMI-SUPERVISED	UBT [22]	ICLR 2021	52.40	73.64	81.52	83.17
	DENT [27]	ECCV 2022	55.81	74.58	82.43	84.32
	DSL [28]	CVPR 2022	54.27	73.21	79.31	81.62
	ARSL [29]	CVPR 2023	54.73	74.74	82.88	84.55
	VC [30]	T-PAMI 2024	54.50	72.50	80.81	82.60
	REMOTEDPL (+△)	~ ~	**58.60** (+5.00%)	**76.13** (+1.86%)	**83.05** (+0.21%)	**84.70** (+0.18%)

THE - DENOTES NO RESULTS RELEASED.

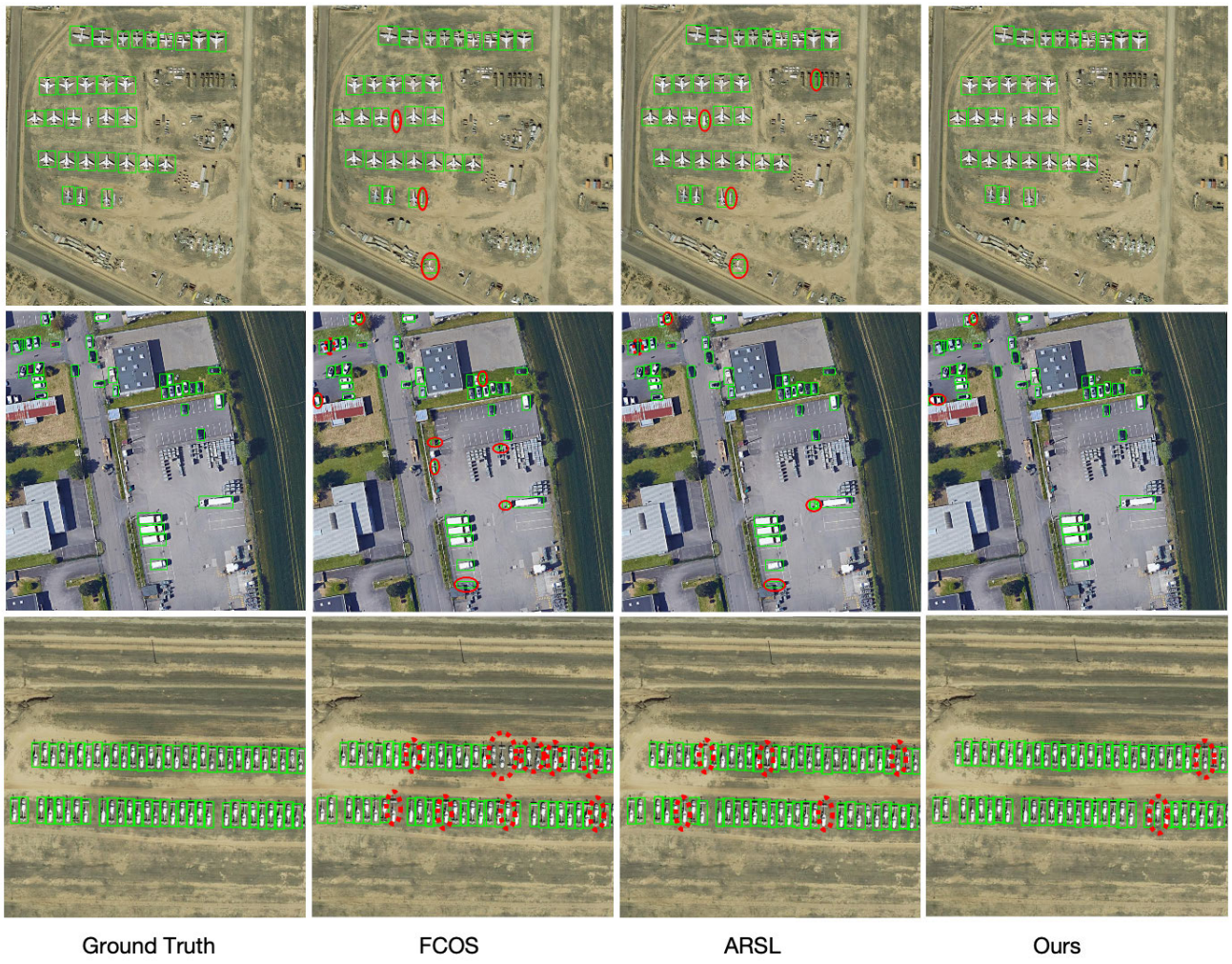

Fig. 7. Visual samples from DOTA-v1.0 at 10% protocol results, presented from left to right as ground truth, the supervised baseline (FCOS) [20], ARSL [29], and our method RemoteDPL. Under the same test settings, RemoteDPL exhibits comprehensive object recognition across various scales in remote sensing datasets and achieves more precise identification in dense scenarios, underscoring its effectiveness. The red dashed and red solid circles represent false negative and false positive, respectively. From the results, it can be seen that our proposed method has significantly fewer error detections compared with the baselines.

and the results are presented in Table I. Specifically, on the validation set, our approach attains mAP scores of 58.60, 76.13, 83.05, and 84.70 for settings with 1%, 5%, 10%, and 100% labeled data, respectively. This represents an improvement over our supervised baseline (i.e., FCOS [20]) by +10.67, +6.82, and +3.48 and-(no results released for FCOS with 100% labeled data) mAP in each respective setting. In addition, our method outperforms ARSL [29], a competitive dense detection method, by +3.87, +1.39, +0.17, and +0.15 in mAP across these proportions of labeled data.

Fig. 7 presents the qualitative results of our method compared with our supervised baseline [20] and ARSL [29], all trained with 10% labeled data. Leveraging the three proposed modules, our model effectively uncovers potential pseudo-labeled boxes, significantly enhancing the detection quality.

TABLE II

EXPERIMENTAL RESULTS ON NWPU VHR-10 UNDER THE PARTIALLY/FULLY LABELED DATA SETTING. EXPERIMENTS ARE CONDUCTED ON 30%, 40%, 50%, AND 100% LABELED DATA SETTINGS. (·) DENOTES THE IMPROVEMENT COMPARED WITH THE OPTIMAL BASELINE

MODEL CATEGORY	METHODS	VENUE	30%	40%	50%	100%
SUPERVISED	FASTER R-CNN [8]	NEURIPS 2016	8.70	73.42	76.62	-
	FCOS [20]	ICCV 2019	65.15	70.16	74.27	-
SEMI-SUPERVISED	UBT [22]	ICLR 2021	70.21	80.78	83.60	86.90
	DENT [27]	ECCV 2022	72.48	82.04	85.37	87.20
	DSL [28]	CVPR 2022	71.07	80.55	82.16	84.20
	ARSL [29]	CVPR 2023	73.22	82.37	85.91	87.62
	VC [30]	T-PAMI 2024	72.66	77.02	82.40	86.22
	REMOTEDPL (+△)	~ ~ ~ ~	**75.74** (+3.44%)	**83.28** (+1.10%)	**87.30** (+1.62%)	**88.19** (+0.65%)

THE - DENOTES NO RESULTS RELEASED.

TABLE III

EFFECTIVENESS OF MULTISCALE JOINT TRAINING, DENSITY ESTIMATION, AND THE OVERALL MODEL. SETTING I IS A VDPLF

Setting	Multi-Scale Joint Training	Density Estimation	Staged Mining	mAP
I	-	-	-	82.31
II	✓	-	-	82.91
III	-	✓	-	82.34
IV	✓	✓	-	82.95
V	✓	✓	✓	**83.05**

TABLE IV

COMPARISON OF THREE FUSION METHODS

Module	mAP
VDPLF (without Fusion)	82.31
Channel-based Fusion [12] (with Fig. 5 (a))	82.73
Our proposed Fusion 1 (with Fig. 5 (b))	82.84
Our proposed Fusion 2 (with Fig. 5 (c))	**82.91**

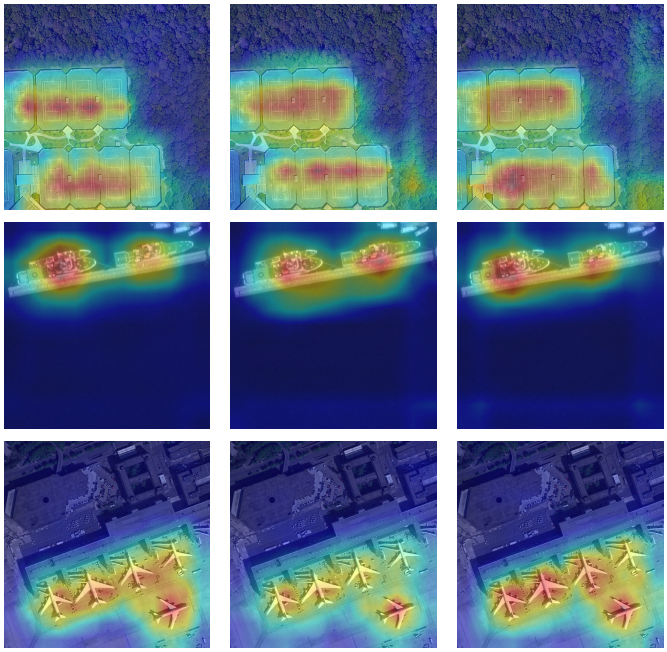

Fig. 8. (Left to right) Visualizations of the features from the fused feature map $P_6^{\times}$ using the three different fusion methods, as illustrated in Fig. 5(a)–(c), respectively.

2) NWPU VHR-10 Dataset: We also evaluate our method on the NWPU VHR-10 dataset under different ratios of labeled data. The results are presented in Table II. On the validation set, our method achieved mAP scores of 75.74, 83.28, and 87.30 with 30%, 40%, 50%, and 100% labeled data, respectively. These results surpass the supervised baseline (i.e., FCOS) by +10.59, +13.12, 13.03, and-(no results released for FCOS with 100% labeled data) mAP. Furthermore, compared with the ARSL [29], which also uses FCOS [20] as the supervised baseline, our approach achieved improvements of +2.52, +0.91, +1.39, and 0.57 mAP at different ratios of labeled data.

Our method demonstrates significant improvements across remote sensing datasets, particularly on the DOTA-v1.0 and NWPU VHR-10 datasets, achieving robust performance at various data ratios. Moreover, comparing contemporary semi-supervised frameworks based on anchor-free [27], [28], [29] and two-stage anchor-based [22], [30] detectors, we find that semi-supervised frameworks with anchor-free detectors achieve better performance on DOTA-v1.0 and NWPU VHR-10.

D. Ablation Study

In this section, we conduct experiments on the DOTA-v1.0 dataset using a 10% data protocol to validate the effectiveness of the proposed module.

1) Effect of Each Component: To investigate the contributions of the three proposed modules, we analyze their roles in enhancing model performance. Without these modules, our approach reduces to a vanilla-dense pseudo-labeling framework (VDPLF). The effects of multiscale joint training, density estimation, and the overall model performance are summarized in Table III. By incorporating multiscale joint training, the mAP improved from 82.31 to 82.91, demonstrating its effectiveness. While the density estimation module did not directly deliver a significant performance boost, this outcome aligns with its intended role: predicting instance density without altering the model's allocation strategy. This module primarily supports the pseudo-label generation phase by providing a basis for secondary mining. To ensure its contribution remains balanced, a weight is assigned to its loss term, minimizing its impact on the overall model. Ultimately, the model achieves an mAP of 83.05, validating the effectiveness of the density estimation module and the staged mining approach. Due to the sparse high-IoU instances in the DOTA-v1.0 dataset, as shown in Fig. 2(b), the number of "pending" samples that can

Fig. 9. Comparison of pseudo-label mining strategies. The first row represents the original single-threshold pseudo-label mining strategy, while the second row illustrates our proposed two-stage pseudo-label mining strategy. Green boxes indicate true positives, orange boxes represent false negatives, and blue boxes denote the mined pseudo-labels. The blue text annotations show the classification score before the "/" and the improvement value after it (please zoomed-in view to better view the annotations in each image).

be effectively recovered through staged pseudo-label mining has decreased, resulting in relatively small performance improvements from staged mining. Although the performance gain (from 82.95 to 83.05) appears modest, the staged mining approach positively contributes by enhancing pseudo-label reliability, leading to a more stable and consistent training process.

2) Influence of Feature Fusion Approaches: In generating hybrid-scale feature maps, we evaluated the impact of three fusion methods. For this analysis, we enhanced the VDPLF by incorporating multiscale joint training and modifying the fusion strategy for hybrid-scale feature maps. Unlike the fusion method in MixTeacher [12], which solely integrates the channel information, we propose two novel fusion approaches that effectively combine spatial and channel information from both the original scale and the downsampled scale.

As presented in Table IV, our proposed fusion methods outperform the approach in MixTeacher. Among them, the fusion method depicted in Fig. 5(c) achieves the best performance and is adopted in our model.

To further illustrate the effectiveness of our approach, Fig. 8 visualizes the performance of the three fusion methods. Under the 10% protocol, our proposed module demonstrates a stronger focus on the target compared with the fusion module used in MixTeacher [12], highlighting its superiority.

3) Selection of the Threshold for the Second Filtering Step: In the second filtering process, the threshold σ_3 plays a critical role. We evaluate the impact of σ_3 on experimental performance, as summarized in Table V. When σ_3 is set to 0.65, the pseudo-labels mined from the "pending" category achieve optimal results. A lower σ_3 results in lower quality pseudo-labels, which negatively impacts the model's detection performance. Conversely, a higher σ_3 reduces the number of mined pseudo-labels, leaving the student model with insufficient pseudo-annotations for effective training. Thus, σ_3 strikes a balance between the quality and quantity of the label.

TABLE V
ABLATION STUDIES ON SECONDARY FILTERING THRESHOLD

Setting	σ_3	mAP
I	0.35	82.47
II	0.45	82.51
III	0.55	82.68
IV	**0.65**	**83.05**
V	0.75	82.56

As illustrated in Fig. 9, our proposed two-stage pseudo-label mining strategy effectively refines the selection of pseudo-labels compared with the original single-threshold approach. The first row of Fig. 9 represents the single-threshold strategy, which may discard valuable pseudo-labels due to the inflexibility of a fixed filtering threshold. In contrast, the second row demonstrates our two-stage mining process, which successfully identifies additional high-confidence pseudo-labels (marked in blue) while maintaining overall detection accuracy. This indicates that our approach not only recovers beneficial pseudo-labels but also mitigates false negatives, ultimately improving the detection performance of the model.

4) Impact of Density Estimation Loss Weights: As shown in Table VI, the model achieves its best performance when β is set to 0.1. This indicates that the teacher model effectively predicts label density scores, enabling the mining of high-quality pseudo-labels without compromising the overall model performance.

TABLE VI
IMPACT OF THE DENSITY ESTIMATION LOSS WEIGHT ON THE OVERALL MODEL

Setting	β	mAP
I	0.5	81.90
II	0.25	82.68
III	**0.1**	**83.05**
IV	0.05	82.58

TABLE VII
COMPARISON OF THREE ARCHITECTURAL PARADIGMS IN CURRENT SSOD

Architectural Paradigms	mAP
Single-branch in Fig. 3 (a)	82.31
Dual-branch in Fig. 3 (b)	82.53
Triple-branch in Fig. 3 (c)	**82.91**

TABLE VIII
PERFORMANCE IMPACT OF THE COMBINED WEIGHT α BETWEEN THE CLASSIFICATION SCORE p AND THE DENSITY SCORE d

Weight α	0.1	0.2	0.3	0.4	**0.5**	0.6	0.7	0.8	0.9
mAP (%)	82.53	82.61	82.70	82.83	**83.05**	82.90	82.72	82.41	82.30

5) Comparison of Three Architectural Paradigms in Current SSOD: As shown in Fig. 3 and Table VII, we further explore the differences among various branch structure paradigms. Specifically, the single-branch paradigm corresponds to VDPLF. For the dual-branch paradigm, to ensure experimental rigor, we adopt the label-level consistency used in PseCo [32], similar to the triple-branch paradigm (i.e., RemoteDPL). The triple-branch paradigm demonstrates significantly stronger performance under the 10% labeled data protocol.

6) Comparison of Model Performance With a Different p and d Weight Distributions: Table VIII shows the impact of different weight distributions between the classification score p and density score d while keeping $\sigma_3 = 0.65$ fixed. The best performance (83.05) is achieved when $p{:}d = 1{:}1$, while shifting the weights in either direction slightly reduces accuracy. This suggests that a balanced contribution from both scores can improve pseudo-label selection. The performance drop is mainly due to fixed σ_3, as changing the weights alters the final confidence score, affecting pseudo-label retention. This highlights the flexibility of our framework, where both weight adjustment and threshold tuning can control pseudo-label quality.

V. CONCLUSION

In this article, we have presented a novel RemoteDPL in remote sensing. Our proposed RemoteDPL addresses two key challenges in remote sensing OD: extreme scale variations and dense object distributions. To address the problem induced by extreme scale variations, our approach leverages the anchor-free framework as baseline, complemented by a fusion module that integrates multiscale spatial and channel information. In addition, we introduce a novel density estimation branch to tackle the complex dense object distributions in remote sensing imagery. By fusing density and classification scores, we implement a two-stage pseudo-label mining strategy that iteratively refines pseudo-label quality, effectively uncovering high-value pseudo-labels even under challenging conditions. Finally, compared with existing methods, our proposed method achieves substantial improvements under limited annotation settings, demonstrating its robustness and adaptability to complex and diverse scenarios encountered in remote sensing applications, with extensive experiments on the two benchmark remote sensing datasets (DOTA-v1.0 and NWPU VHR-10). As a potential future work, the proposed method can be further extended from the following two aspects. First, the strategies for fusing original- and downsampled-scale features could be further improved to obtain better representations. Second, the concept of object density modeling used in our method can be extended to other tasks (e.g., semantic segmentation and general OD), where scenes with dense annotations and drastic variations in object scale are usually encountered.

ACKNOWLEDGMENT

The authors sincerely thank the editor and anonymous reviewers for their helpful comments to further improve this article.

REFERENCES

[1] S. Lin, "Automatic recognition and detection of building targets in urban remote sensing images using an improved regional convolutional neural network algorithm," *Cognit. Comput. Syst.*, vol. 5, no. 2, pp. 132–137, Jun. 2023.
[2] P. Berg, M.-T. Pham, and N. Courty, "Self-supervised learning for scene classification in remote sensing: Current state of the art and perspectives," *Remote Sens.*, vol. 14, no. 16, p. 3995, Aug. 2022.
[3] H. Lang, M. Agrawal, Y. Kim, and D. Sontag, "Co-training improves prompt-based learning for large language models," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2022, pp. 11985–12003.
[4] C. Doersch, "Tutorial on variational autoencoders," 2016, *arXiv:1606.05908*.
[5] R. Labaca-Castro, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 73–76.
[6] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1278–1286.
[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 91–99.
[9] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," 2018, *arXiv:1809.08545*.
[10] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4507–4515.
[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.
[12] L. Liu et al., "MixTeacher: Mining promising labels with mixed scale teacher for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7370–7379.
[13] Y. Ma, S. Lan, W. Ma, X. Yin, and Y. Li, "Dense pseudo-labels based semi-supervised object detection for remote sensing," in *Proc. Int. Joint Conf. Neural Netw.*, 2024, pp. 1–9.
[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[15] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[17] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[18] X. Yin et al., "DLAHSD: Dynamic label adopted in auxiliary head for SAR detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 3434–3438.

[19] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[20] 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. 9627–9636.

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

[22] Y.-C. Liu et al., "Unbiased teacher for semi-supervised object detection," 2021, *arXiv:2102.09480*.

[23] Q. Yang, X. Wei, B. Wang, X.-S. Hua, and L. Zhang, "Interactive self-training with mean teachers for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 5941–5950.

[24] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell. (IJCAI)*, 1981, pp. 674–679.

[25] Q. Zhou, C. Yu, Z. Wang, Q. Qian, and H. Li, "Instant-teaching: An end-to-end semi-supervised object detection framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4081–4090.

[26] Y. Tang, W. Chen, Y. Luo, and Y. Zhang, "Humble teachers teach better students for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3132–3141.

[27] H. Zhou et al., "Dense teacher: Dense pseudo-labels for semi-supervised object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 35–50.

[28] B. Chen, P. Li, X. Chen, B. Wang, L. Zhang, and X.-S. Hua, "Dense learning based semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4805–4814.

[29] C. Liu et al., "Ambiguity-resistant semi-supervised learning for dense object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15579–15588.

[30] C. Chen, J. Han, and K. Debattista, "Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5595–5611, Aug. 2024.

[31] Q. Guo, Y. Mu, J. Chen, T. Wang, Y. Yu, and P. Luo, "Scale-equivalent distillation for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14502–14511.

[32] G. Li, X. Li, Y. Wang, Y. Wu, D. Liang, and S. Zhang, "PseCo: Pseudo labeling and consistency training for semi-supervised object detection," in *Proc. 17th Eur. Conf. Comput. Vis.* Berlin, Germany: Springer-Verlag, 2022, pp. 457–472.

[33] G. Liu, F. Zhang, T. Pan, and B. Wang, "Low-confidence samples mining for semi-supervised object detection," 2023, *arXiv:2306.16201*.

[34] G. Chen, L. Liu, W. Hu, and Z. Pan, "Semi-supervised object detection in remote sensing images using generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2503–2506.

[35] B. Zhang, Z. Wang, and B. Du, "Boosting semi-supervised object detection in remote sensing images with active teaching," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[36] J. Hun Wang et al., "Weakly-semi-supervised object detection in remotely sensed imagery," 2023, *arXiv:2311.17449*.

[37] R. Fu et al., "S^2O-det: A semisupervised oriented object detection network for remote sensing images," *IEEE Trans. Ind. Informat.*, vol. 20, no. 9, pp. 11285–11294, Sep. 2024.

[38] R. Fu, C. Chen, S. Yan, X. Wang, and H. Chen, "Consistency-based semi-supervised learning for oriented object detection," *Knowl.-Based Syst.*, vol. 304, Nov. 2024, Art. no. 112534.

[39] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017, pp. 1195–1204.

[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[41] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.

[42] 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.*, Jul. 2017, pp. 2117–2125.

[43] 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.

[44] 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.

[45] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.