Dual teacher: improving the reliability of pseudo labels for semi-supervised oriented object detection.

FANG, Z., REN, J., ZHENG, J., CHEN, R. and ZHAO, H.

2024

Dual Teacher: Improving the Reliability of Pseudo Labels for Semi-Supervised Oriented Object Detection

Zhenyu Fang, Jinchang Ren, Jiangbin Zheng, Rongjun Chen and Huimin Zhao

***Abstract*—Oriented object detection in remote sensing is a critical task for accurately location and measurement of the interested targets. Despite of its success in object detection, deep learning-based detectors rely heavily on extensive data annotation. However, variations in object appearance significantly increase the difficulty and the cost of creating large-scale annotated datasets. Semi-supervised learning aims to utilize unlabeled data to enhance object detectors. Among these, pseudo-label-based methods have shown promising results recently. Nonetheless, as training progresses, the accumulation of errors in pseudo-labels leads to prediction bias without corrections. To tackle this particular challenge, we present a semi-supervised learning pipeline, named "Dual Teacher", for improving the reliability of pseudo labels in the semi-supervised oriented object detection. Firstly, to mitigate the bias caused by limited annotated data, a global burn-in strategy is introduced at the beginning of training, which guides the student detector to learn the feature extraction on a global scale. Additionally, an online bounding box correction module is proposed to decrease the occurrence of mislabeled instances and enhance the reliability of detection. These improvements are facilitated by an additional detector, instead of a single teacher model in the teacher-student architecture. Dual Teacher reduces the dependency on the quality of pseudo-labels related to the model complexity, and combines the strengths of both the two-stage and one-stage detectors. With only 20% labeled data, Dual Teacher outperforms fully supervised R-FCOS, YOLOX-s and R-RCNN by up to 2% on both DOTA and SODA-A datasets. This reveals its potential in reducing labor-intensive tasks and enhancing robustness against environmental interference and noisy labels. The code is available at https://github.com/ZYFFF-CV/DualTeacher-semisup.git.**

***Index Terms*—Oriented Object Detection, Semi-Supervised Learning, Pseudo Label, Dual Teacher, Consistency Learning.**

I. Introduction

ORIENTED object detection is crucial in optical remote sensing imagery, pivotal for applications ranging from environmental monitoring to urban planning [1]. Unlike conventional object detection, which often occurs under controlled conditions with objects appearing at similar scales and orientations, remote sensing imagery introduces distinct challenges [2], [3]. Captured from aerial viewpoints, these images display objects at diverse scales and orientations, influenced by sensor altitude and angle. This necessitates the development of specialized detection algorithms capable of handling the significant variability in object appearance and environmental conditions effectively.

Existing supervised learning approaches depend heavily on large, well-annotated datasets [4], [5], which are often costly and labor-intensive to create [6], particularly due to the expansive and detailed nature of remote sensing data. In contrast, semi-supervised learning, which utilizes a small amount of labeled data along with a larger pool of unlabeled data, offers an effective solution. This approach significantly reduces the reliance on extensively labeled datasets while improving learning accuracy and model robustness. Semi-supervised learning not only overcomes the challenge of limited labelling data but also adapts effectively to the complex and varying characteristics of remote sensing imagery, thus proving to be invaluable for advancing object detection technologies in this field.

Pseudo-label-based approaches are widely used in semi-supervised learning [7], [8], [9], [10]. Initially, the referred "teacher model" [11], [12], is trained on the available labeled data. As illustrated in Fig. 1, this teacher model then applies the learned knowledge to predict labels for unlabeled images within the dataset. These predictions, known as pseudo-labels, are selected based on a confidence threshold, where only labels exceeding a specific confidence level, either manually assigned [9], [10] or ranked in the top-N [11], [12], are retained. These high-confidence pseudo-labels, presumed accurate, are utilized to augment the original training dataset. Subsequently, the expanded annotations are used to train another model, denoted as the "student model". To prevent the prediction failure, the teacher model is updated more gradually than the student model. To improve the quality of the pseudo labels, the learned knowledge of the teacher model is further leveraged, using various strategies such as feature map distillation [13], multi-scale learning [12] and weakly supervised learning [14]. With the support of high-quality pseudo-labels, these methods have shown promising results in detecting objects from general scenes.

In the context of remote sensing object detection, however,

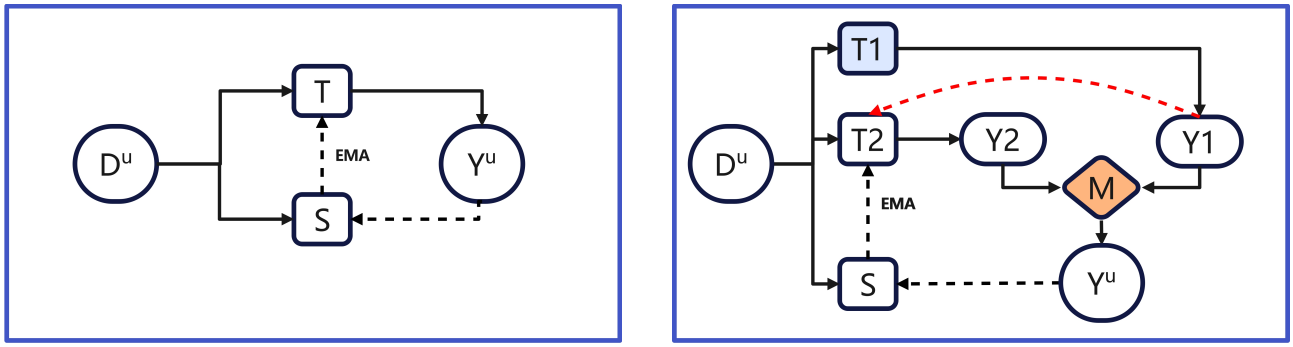
This work was supported in part by the National Natural Science Foundation of China under Grant 62202385, in part by the Fundamental Research Fund for the Central Universities under Grant G2021KY05103, and in part by the Basic Research Programs of Taicang under Grant TC2022JC21. (Corresponding author: Jinchang Ren, jinchang.ren@ieee.org).

Z. Fang, and J. Zheng are with the School of Computer Software, Northwestern Polytechnical University (NPU), Xi'an, China. E-mail: zhenyu.fang@nwpu.edu.cn

Z. Fang is also with Yangtze River Delta Research Institute of NPU, Taicang, China

J. Ren, R. Chen and H. Zhao are with School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China

Prof. Ren is also with National Subsea Centre, Robert Gordon University, Aberdeen, UK



(a) (b)

Fig. 1: Comparison of schemes for processing unlabeled images between (a) existing pseudo-label-based semi-supervised learning methods and (b) the proposed Dual Teacher method. "D^u" represents the unlabeled images and "Y^u" indicates the pseudo labels. Forward propagation and backward propagation are denoted by solid lines and dashed lines, respectively. In (a), the teacher model and the student model adopt the same architecture, and the teacher model's weights are updated via the student model using exponential moving average decay (EMA). To mitigate noise interference from the pseudo labels, the proposed Dual Teacher employs an additional teacher model with a heterogeneous architecture (denoted as "$T1$"), which allows the senior model (denoted as "$T2$") to learn the global-wise dataset distribution initially (represented by the red dashed line). Concurrently, an online bounding box correction strategy (denoted by "M") is implemented to improve the quality of pseudo labels.

existing pseudo-label-based methods may suffer from several limitations.

i. Unreliable pseudo-labels: the variability of object appearance due to different environmental and imaging conditions may increase the distribution gap between the labeled and the unlabeled sets, causing the teacher model to struggle in making high-confidence predictions. This hinders the weight optimization of the student model;
ii. Error accumulation: existing methods ensure the quality of pseudo-labels using a confidence threshold. However, this approach also excludes many potential foreground predictions. As training progresses, errors accumulate from the student model to the teacher model, remaining uncorrected;
iii. Unbalanced efficacy and efficiency: since the teacher model's weights are updated from the student model, both models must maintain the same architecture. Given the significant imbalance between the foreground and background instances in remote sensing, two-stage detectors typically outperform one-stage detectors in efficacy. However, the inference efficiency of two-stage detectors is less optimal.

To mitigate these challenges, a novel semi-supervised oriented object detection method in the remote sensing pipeline, termed "Dual Teacher", is proposed. As illustrated in Fig. 1(b), an off-the-shelf heterogeneous teacher model, named the "supervisor model", is employed from the start to the end of the training process to assist in noise reduction caused by pseudo labels. Although the supervisor model is a two-stage detector, it does not participate in the weight update process, which decouples the quality of pseudo-labels from the model architecture. Notably, the supervisor model is pre-trained using another semi-supervised learning method on the same dataset, i.e. Soft Teacher [10], thus adding no additional annotation burden. The major contributions of the proposed Dual Teacher are summarized as follows:

i. A global burn-in method is introduced at the beginning of training to allow the detector to learn feature extraction on a "dataset-wise" basis rather than from a "local-wise" annotated subset [10], [11]. This helps make the teacher model, aligned with the "senior model" in Dual Teacher, more robust to the variability of object appearances, especially when the annotation set is biased and limited;
ii. An online bounding box (bbox) correction strategy is implemented to enhance the quality of the pseudo-labels and reduce the false negative instances. This is achieved by integrating the predictions from the two teacher models with heterogeneous architectures. Different from the existing methods, where the quality of pseudo bbox can only be refined through multi-task learning, our strategy is more reliable, and is robust to the ambiguity of the teacher model;
iii. Experimental results on the DOTA [15] and SODA-A [16] datasets demonstrate that Dual Teacher, using a single-stage detector, outperforms the peers and exceeds the performance of fully supervised learning even with only 20% of annotated data.

The remainder of this paper is structured as follows. Section II reviews related works, and Section III details the proposed method. Experimental results are presented in Section IV, and the conclusion is drawn in Section V.

II. Related Works

A. Semi-supervised Object Detection

Semi-supervised learning (SSL) methods leverage the potential of unlabeled images when annotated images are scarce.

Broadly, SSL approaches fall into two categories: consistency-based and pseudo-label-based methods.

Consistency-based methods [17], [18], [19], [20], [21], [22] seek to maintain uniformity across manually introduced perturbations. A regularization loss quantifies the discrepancy among these perturbations, aiming to minimizing these discrepancies. Perturbations can be applied to the model [17], the images [18], or even the data distribution [19]. Tarvainen et al. [7] introduced a teacher-student paradigm known as "Mean Teacher," where consistency is enforced between the outputs of the student and teacher networks. The teacher network's weights are updated from the student's through an exponential moving average (EMA) strategy.

Pseudo-label-based methods [23], [24], [25], [26], [27] generate annotations for unlabeled images using the knowledge gained from the labeled data. These pseudo-labeled images are then utilized to further refine the model. This approach has been widely adopted in recent SSL-based object detection methods. Specifically, STAC [26] is built on the Mean Teacher workflow. Initially, an image undergoes two types of augmentations: weak and strong, where strong augmentation introduces greater variations compared to weak augmentation. Images subjected to weak and strong augmentations are fed into the student and teacher networks, respectively. The output from the teacher network serves as the pseudo-label for training the student network. Subsequently, Unbiased Teacher [8] replaces the cross-entropy loss with a focal loss [28] to address class imbalance. In its second iteration, Unbiased Teacher v2 [9] explores the potential of anchor-free detectors in SSL contexts. Soft Teacher [10] assesses the reliability of pseudo-labels based on the teacher network's scoring and the stability of bounding boxes against jitter, significantly enhancing pseudo-label quality. Dense Teacher [13] advances this by exploiting dense pixel-level pseudo-labels to analyze background features, as opposed to solely focusing on instance-level data.

More recently, SOOD [11] has been introduced to address the challenge of detecting oriented objects in remote sensing, employing a rotation-aware adaptive weighting (RAW) loss and a global consistency (GC) loss to mitigate background interference. Previous studies have significantly contributed to enhancing pseudo-label quality, typically assessed internally since the teacher network is essentially a time-shifted version of the student network. As training progresses, errors can accumulate, potentially trapping the student in suboptimal performance. In contrast, our proposed method introduces two teacher models with distinct weights and architectures to mitigate error accumulation effects, which is detailed in Section III.

B. Oriented Object Detection

Similar to general-purpose object detection methods, oriented object detectors used in remote sensing can be broadly classified into two-stage and one-stage detectors.

Two-stage detectors [29], [30] identify objects using two subnetworks. Initially, a backbone [31], [32], integrated with a Feature Pyramid Network (FPN) [30], extracts multi-scale feature maps. A Region Proposal Network (RPN) then identifies potential foreground instances, also known as regions of interest (ROIs). Local features are extracted from the feature map using ROI pooling or ROI align layers, followed by dual-stream convolutional networks that categorize and refine bounding box predictions. Considering the unique needs of remote sensing, oriented bounding boxes [33] and rotated ROI align layers [34] are also employed to handle the rotation angles of objects. Given that images in remote sensing are typically larger than 1000 pixels with objects smaller than 50 pixels [15], [16], class imbalance between the foreground and background presents a significant challenge. Two-stage detectors address this issue by resampling the ratio of foreground to background, albeit at the cost of reduced computational efficiency.

In contrast, one-stage detectors employ an FPN-based backbone for feature extraction with a single subnetwork for detection. Without the need for resampling ratios of the foreground and background, the focal loss [28] is implemented to recalibrate the contribution to gradient updates based on detection difficulty. This adjustment has propelled the RetinaNet to achieve commendable performance compared to the two-stage detectors, with higher computational efficiency. Further refinements in angle prediction employ a Gaussian-based reweighting scheme [35], [36], [37]. To enhance inference speed, anchor-free methods [38], [39], [40] have been developed to predict object dimensions without predefined human priors, leading to one-stage detectors' predominance in remote sensing.

Additionally, we notice that dual teacher has been used in two published works [41], [42], though the novelties are quite discriminate as analyzed in detail as follows. Zheng et al. [41] utilize a two-teacher model on the mutual interference between the optical and SAR for supervised ship detection. Xin et al. [42] propose a semi-supervised semantic segmentation framework, integrating two teacher models on the consistency regularization learning and contrastive learning (CL), respectively. Thus, the two teacher models utilized in these methods are proposed for multi-task learning. On the contrast, our method adopts the knowledge inheritance between two teacher models and the supervisor model in our Dual Teacher is not updated during the training. This allows the pseudo-labels can be well corrected to alleviate the interference on the annotation error.

However, when applied to semi-supervised learning (SSL), one-stage detectors may underperform compared to two-stage models due to the invalid assumption of completely reliable annotations underlying the focal loss, particularly in remote sensing. This may severely affect the training performance, including numerous false negative samples and few false positives, where the situation can be worsened when directly applying focal loss. To harness the strengths of both detector types, the proposed method integrates their architectures into the training process, as detailed in subsequent sections.

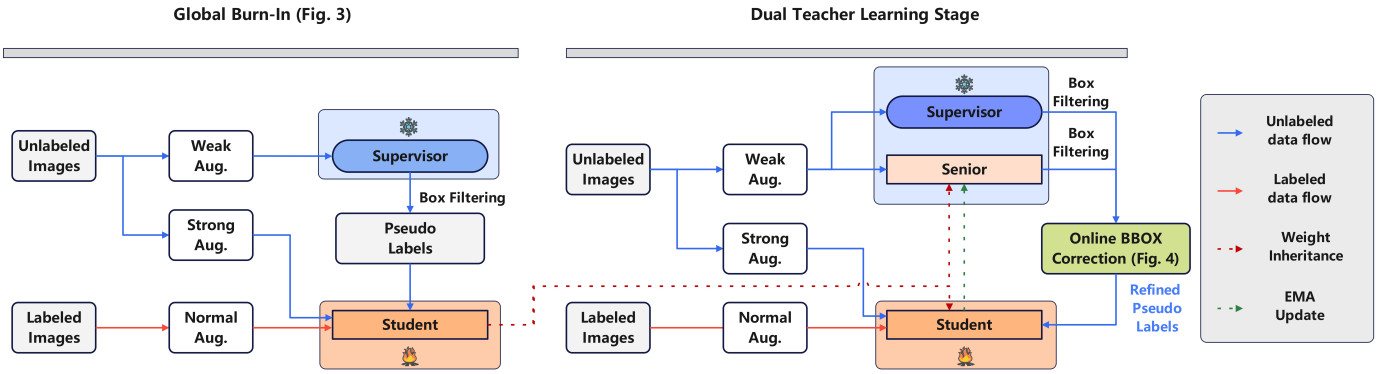


Fig. 2: The pipeline of the proposed Dual Teacher method consists of two stages: Global Burn-in and Dual Teacher Learning. Initially, as detailed in Fig. 3, a conventional two-stage detector is utilized to generate pseudo labels from unlabeled images, during which the student model optimizes the weights using both labeled and pseudo-labeled subsets. Subsequently, the weights of the student model are transferred to the senior model. In the Dual Teacher Learning stage, pseudo labels are collaboratively refined by the Senior model and the Supervisor model, facilitated by the newly proposed online bounding box correction module, where the process is detailed in Fig. 4. The Senior model's weights are updated using an Exponential Moving Average (EMA) strategy from the student model.

III. Proposed Method

A. Preliminary and Motivation

The overall pipeline adheres to pseudo-label-based methods. Given a dataset of labeled images $D^s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ and a set of unlabeled images $D^u = \{x_i^u\}_{i=1}^{N_u}$, where N_s and N_u are the numbers of labeled images and unlabeled images, respectively. These methods train detectors using semi-supervised learning. The teacher-student architecture is employed to learn feature extraction from the unlabeled images, where the teacher network generates pseudo labels. Diverse augmentations are crucial: weak augmentation is applied to images processed by the teacher model, whereas strong augmentation, involving more intensive preprocessing steps, is used for images input to the student model. For labeled images, standard augmentation procedures suffice [43], as they are employed solely for supervised learning.

The loss function for the student model encompasses both supervised learning loss and pseudo-label learning loss, which can be formulated as:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_s\left(P^s, Y^s\right) + \mathcal{L}_u\left(P^u, Y^u\right) \\ &= \sum_i \mathcal{L}_s\left(p_i^s, y_i^s\right) + \lambda_u \sum_i \mathcal{L}_u\left(p_i^u, y_i^u\right)\end{aligned} \tag{1}$$

where p_i^s and p_i^u are the predictions for labelled images and unlabeled images, respectively. The contribution of the pseudo-label loss is adjusted by a weight parameter λ_u.

After the parameter upgrade of the student model, the weight of Teacher model is upgraded based on Exponential Moving Average (EMA) as introduced in [7], defined by:

$$\theta_{\text{teacher}}^{\text{t}} = m\theta_{\text{teacher}}^{\text{t}-1} + (1-m)\theta_{\text{student}}^{\text{t}} \tag{2}$$

where $\theta_{\text{teacher}}^t$ and $\theta_{\text{student}}^t$ represent the weights of the teacher and student models at the training step t, respectively. The parameter m denotes the momentum, which adjusts the intensity of updates. With limited training data, the semi-supervised learning based detectors can achieve a comparable mean Average Precision (mAP). However, there remains a performance gap when compared with the fully-supervised-learning-based detectors [44], [45], [46], [47].

Based on the preliminary results, we present the motivation for the proposed method. Existing pseudo-label-based methods derive labels from the teacher model; however, they can lead to error accumulation. We deduce that the errors originate from both the biased feature estimation at the onset of training and the Exponential Moving Average (EMA) based weight update workflow. To elucidate further, consider a dataset in a remote sensing scenario denoted as $D_l = \{x_i^l, y_i^l\}_{i=1}^{N_l}$, where $N_l = N_s + N_u$. The weights of a detector trained with D_l and D_s are represented by θ_l and θ_s, respectively. Given that D_l includes unseen samples from D_s, θ_l is inherently more robust. This relationship can be mathematically expressed as follows:

$$\theta_l = \theta_s + \Delta\theta \tag{3}$$

where $\Delta\theta$ represents the weight offset. As $\Delta\theta$ increases, the offset also enlarges correspondingly. In the extreme case, when D_l encompasses all samples from the testing scenarios, with an adequate distribution, θ_l might be sub-optimal but is still considered a better solution than θ_s.

Considering that multiplication (division), addition (subtraction), and the ReLU function are the most frequently used operations in many state-of-the-art detectors [48], [49], [50], the detector $f(x, \theta_l)$ can be expressed as follows:

$$\begin{aligned}f\left(x, \theta_l\right) &= f\left(x, \theta_s + \Delta\theta\right) \\ &= \left(w_s + \Delta w\right) x + \left(b_s + \Delta b\right) \\ &= f\left(x, \theta_s\right) + f(x, \Delta\theta)\end{aligned} \tag{4}$$

Here, w_s and b_s denote the weight and bias, respectively. For simplicity, Batch Normalization (BN) and ReLU are omitted. As observed, with limited training data, a prediction error is inevitable. At the onset of training, detectors are

primarily capable of learning feature extraction from labeled images, a phase often referred to as "burn-in". As discussed in Eq. 3, both the two-stage burn-in [8], [9] and end-to-end learning approaches [10] can lead to prediction bias if the parameters are not properly calibrated.

As training progresses, the detector increasingly relies on the quality pseudo-labels predicted by the teacher model. Since the weights of the teacher model are updated based on the student model's outputs, errors from the student model gradually accumulate in the teacher model. Existing methods have proposed bounding box (bbox) filtering strategies, such as "listen-to-student" [9] and bbox jittering [10], to eliminate low-quality bboxes. However, these strategies still rely on features referenced from either the teacher or the student model, which diminishes the confidence in quality estimation.

B. The Overall Pipeline

Motivated by the limitations discussed previously, we propose a semi-supervised learning method named "Dual Teacher" for remote sensing images. As depicted in Fig. 2, our approach diverges from prior work by employing two teacher in the proposed pipeline to maximize the use of unlabeled images at the outset of training and enhance the reliability of pseudo boxes. These models are referred to as the "supervisor model" and "senior model," respectively. Initially, we introduce a global burn-in strategy to mitigate biased learning due to limited labels. This involves using an off-the-shelf detector (supervisor model) as an annotator for unlabeled data, coupled with a lateral-learning strategy to modulate the weight update pace of the senior model, details of which are provided in Section III-C. Subsequently, we introduce an online bounding box (bbox) correction module to minimize noise from annotation errors in pseudo labels. This module allows for the correction of pseudo labels using two distinct models, and the student model's weights are updated based on "high-quality" pseudo labels. This approach is further elaborated in Section III-C. Employing a single-stage detector, our proposed method has demonstrated superiority over existing two-stage detectors in the field of remote sensing, particularly with limited annotated images.

There are several differences between the supervisor model and the senior model. First, for gradient upgrade, the weight of the senior model is EMA updated based on the weights of the student model, while the weight of the supervisor model is frozen during the training process. In other words, the senior model has no connection to the supervisor model during the gradient upgrade. Second, the architectures of those two models are different. The senior model has the same architecture as the student model, i.e., a one-stage detector, whilst the supervisor model is a pretrained two-stage detector, which is not trained through training. As the weight of supervisor model is not updated, the pseudo label noise will not be accumulated during training. Additionally, using two teacher models with diverse architectures will inherit the advantages from both models.

Boosted by the two proposed modules, the framework of the proposed Dual Teacher is summarized in Alg. 1. The senior model is adopted for evaluation, as it is EMA updated from the student model, where the weight of the senior model is more robust to the gradient oscillation caused by the pseudo noise within a certain batch [7]. A similar strategy is also applied in the YOLO series [51], where the learned weights are firstly EMA updated in a "buffer zone" and will be cloned to the model after the completion of an epoch.

Algorithm 1 The proposed Dual teacher learning paradigm

Input: Supervisor model $g(x, \theta_{sup})$, senior model $f(x, \theta_{sen})$, student model $f(x, \theta_{st})$.
Labelled image set $D^s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ and unlabeled image set $D^u = \{x_i^u\}_{i=1}^{N_u}$
Output: Well trained senior model $f(x, \theta_{sen})$
Shuffle the dataset D^s and D^u, respectively.
$t \leftarrow 0$
while $t \leq T_{max}$ **do**
  $X \leftarrow \{X^s, X^u\}$, $\forall X^s \subset D^s$ and $\forall X^u \subset D^u$ ▷ Fetch mini-batch with a ratio of 1:4
  $X^u_{strong} \leftarrow$ Strong augmentation on X^u
  $X^u_{weak} \leftarrow$ Weak augmentation on X^u
  Augment X^s with commonly used method
  $Y^u \leftarrow g(X^u_{weak}, \theta_{sup})$
  if $mAP_{student} \geq mAP_{supervisor}$ **then**
    if θ_{se} is not initialized **then** ▷ Upgrade Senior model parameters
      $\theta_{sen} \leftarrow \theta_{st}$
    else
      $\theta_{sen} \leftarrow m\theta_{sen} + (1-m)\theta_{st}$
    end if
    $Y^u_{sen} \leftarrow f(X^u_{weak}, \theta_{sen})$
    $Y^u \leftarrow NMS(Y^u, Y^u_{sen}, th)$ ▷ Class-wise NMS
  end if
  $P^s \leftarrow f(X^s, \theta_{st})$ ▷ Calculate Student model predictions
  $P^u \leftarrow f(X^u_{strong}, \theta_{st})$
  Calculate loss as in Eq. 1, and get gradient $\Delta\theta_{st}$
  Upgrade $\theta_{st} \leftarrow \theta_{st} - \eta\Delta\theta_{st}$ ▷ η is the learning rate
  $t \leftarrow t + 1$
end while

C. Global Burn-in Strategy

Given a sufficient number of training images, deep learning models can significantly enhance their feature extraction capabilities, particularly in the field of remote sensing where the scenes may vary considerably between images. To effectively utilize the unlabeled data, we propose a global burn-in strategy. In this strategy, a pretrained detector, referred to as the "supervisor" within our dual teacher method, is employed to annotate unlabeled data. Training the supervisor model with additional annotated data would necessitate extra manual labor, which is contrary to the goals of semi-supervised learning. Consequently, as seen in Fig. 3, the supervisor model is pretrained using a state-of-the-art semi-supervised learning method, specifically the Soft Teacher model [10], where high confidence pseudo labels are selected for unlabeled images.

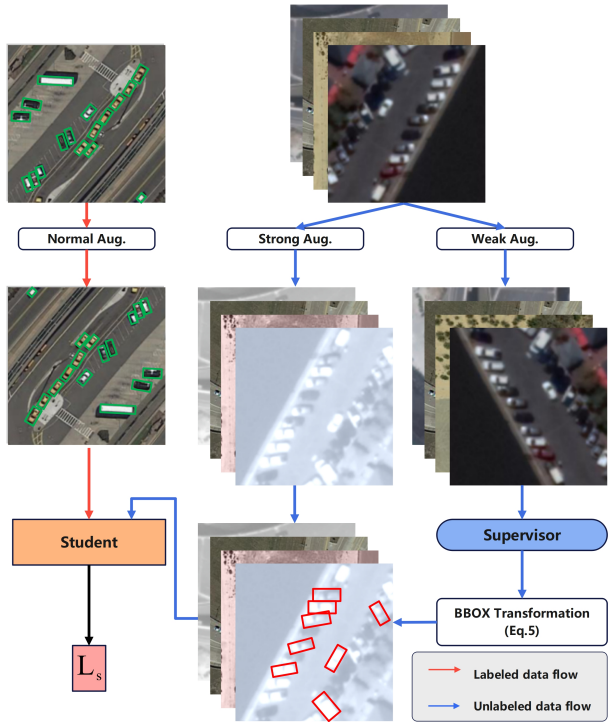


Fig. 3: Flowchart of the proposed global burn-in strategy (GBI). Different from the existing work, the student model is trained from both labeled and unlabeled images, where the latter use pseudo labels generated from the learned Teacher Model in a supervised way.

In this context, no additional annotation is needed during the global burn-in stage. Moreover, the supervisor model does not update its parameters throughout the training process. Thus, during the global burn-in phase, the student model learns in a manner akin to the supervised learning.

Specifically, due to the diversity of the two augmentation pipelines, the bounding boxes (bboxes) predicted by the supervisor model cannot be directly applied to training. We adopt a methodology similar to that used in the Soft Teacher approach, which involves aligning the augmentations for the predicted bboxes as described below:

$$\mathfrak{B}_{sp} = A_{st} A_{we}^{-1} \mathfrak{B}_{sp}^{ori} \tag{5}$$

where $\mathfrak{B}_{\text{sp}}^{\text{ori}}$ represents the set of oriented bounding boxes (bboxes) predicted by the supervisor model in "XYXY" format. The matrices A_{st} and A_{we} denote the transformation matrices for strong and weak augmentations, respectively. This alignment approach will also be applied to the predicted bboxes of the senior model, facilitating the online bounding box correction method.

Existing consistency-based semi-supervised learning method needs to upgrade the weights of the senior model at the beginning of training, though the student model has a low prediction accuracy. This is because their pseudo labels used in the existing methods are only sourced from one teacher model (aligned to the senior model in the proposed Dual Teacher method). If their teacher model remains unchanged, the student model may collapse.

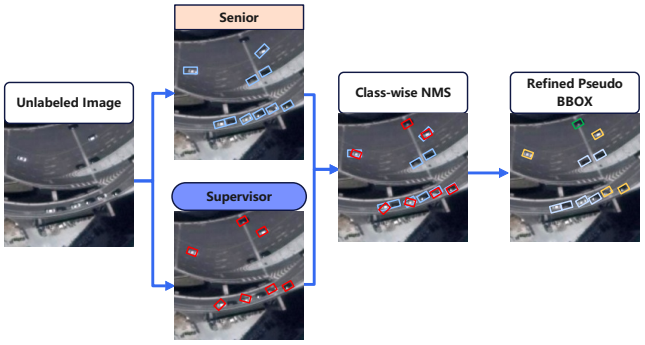


Fig. 4: Illustration of the proposed online BBOX correction strategy. BBOXes predicted by the Senior model are annotated in light blue, while the BBOXes predicted by the supervisor model are annotated in red. As seen, with the aid of supervisor model, the missing annotation can be re-annotated (highlighted in greed) and the low-quality bbox can be refined (highlighted in yellow).

As opposed to these pipelines, the proposed lateral-learning strategy does not inherit the weights from the student model until the student is well learnt. This will allow the senior model more robust to the pseudo-label noise, predicted from the supervisor model, and can further improve the quality of pseudo labels in the following online bbox correction process.

As mentioned above, there exists another teacher model, referred to as the "Senior". Recall the Exponential Moving Average (EMA)-based weight update strategy discussed in Eq. 2, where typically, the teacher model updates its weights at the beginning of the training. In this paper, however, updates are made only when the student model outperforms the supervisor model. Consequently, the momentum parameter m will be assigned as follows:

$$m = \begin{cases} 0, & \text{if } mAP_{\text{student}} \geq mAP_{\text{supervisor}} \\ 0.999, & \text{else} \end{cases} \tag{6}$$

where the value of m is adopted from [10], [11].

To mitigate the noise associated with pseudo labels, we filter out low-quality labels using a high threshold. It is evident that a high threshold may lead to a significant number of false negatives, i.e., as missing annotated objects. It is particularly critical in the remote sensing field where small objects predominate, hence our approach is essential for stabilizing the training process. Nonetheless, missing annotations can be re-annotated through the proposed online bounding box (bbox) correction method, the details of which will be elucidated in the following section. Different from Soft Teacher, the reliability of pseudo bboxes is not estimated via background prediction scores, this is because the single-stage methods can only predict low score on foreground instances, under insufficient manual labels. This phenomenon is also reported in our experiments. Thus, applying background score into the reliability estimation may cause more potential foreground bboxes discarded.

D. Online BBOX Correction

Although the global burn-in enhances the student model, issues of missing and inaccurate annotations persist. To address these errors, an online bounding box (bbox) correction scheme is proposed, employing both teacher models to annotate unlabeled images. As seen in Fig. 4, following the global burn-in stage in Fig. 3, bboxes predicted by the senior model (denoted by $\mathfrak{B}_{\text{se}}$) and the supervisor model (denoted by $\mathfrak{B}_{\text{sp}}$) are fused through Non-Maximum Suppression (NMS), to exclude low confidence duplicated pseudo bbox predicted from both $\mathfrak{B}_{\text{se}}$ and $\mathfrak{B}_{\text{sp}}$, The remaining bboxes format the unsupervised labels (denoted by $\mathfrak{B}_t$). Note that the NMS is conducted solely within the same class. This process can be expressed as follows:

$$\mathfrak{B}_t = \left\{ c \in (1, 2, \ldots, C) \mid NMS\left(\mathfrak{B}^c_{se}, \mathfrak{B}^c_{sp}\right) \right\} \tag{7}$$

Note that Non-Maximum Suppression (NMS) is conducted solely within the same class. Before the online bounding box (bbox) correction, it is crucial to eliminate unreliable bboxes from each teacher model. For the supervisor model, the filtering rule remains consistent with that used during the global burn-in stage. Similarly, bboxes from the senior model are also filtered using "threshold truncation", where bboxes with confidence levels below a specified threshold are discarded. Although multiple bbox merging methods exist, such as Soft NMS [52] and other variants [53], the vanilla NMS proves to be the most effective in terms of mean Average Precision (mAP). This will be detailed in Section IV.

E. Loss Functions

As described in Eq. 1, the loss function is composed of two components: supervised loss and unsupervised loss. Reflecting the architecture characteristics of both the senior and the student models, the same format is employed for both loss functions in this paper. Specifically, the classification loss utilizes the focal loss [28], and the regression loss employs the Intersection Over Union (IOU) loss [38]. Notably, the pseudo label set Y^u consists of bounding box coordinates with category IDs, identical to the manual label set Y^s. We hypothesize that this uniform format is advantageous for reducing the noise associated with pseudo labels, which will be discussed in more detail in Section IV.

IV. Experimental Results

A. Dataset and Evaluation Protocol

Two benchmark datasets are utilized as benchmarks [15], [16], [11], [34] in this paper.

DOTA [15]. DOTA is a large-scale dataset used in remote sensing for object detection tasks. It comprises 2806 large aerial images and 402,089 annotated oriented objects. The dataset is segmented into three subsets: training, validation, and testing, containing 1411, 458, and 937 images respectively. Objects within these images are classified into 16 categories: 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), Helicopter (HC), and Container crane (CC). Following prior research [11], annotations from DOTA v1.5 are employed, which include additional instances of extremely small objects (less than 10 pixels). Due to the large size of the raw images, following established methodologies [15], [11], [37], the original images are cropped into 1024 $\times$ 1024 patches with a stride of 824 pixels, ensuring a 200-pixel overlap between adjacent patches. Since the annotations for the DOTA-v1.5-test subset are not released, results are reported using the validation set.

TABLE I: Number of instances in the DOTA v1.5 Dataset under annotated proportions of 10%, 20% and 30%, respectively. "L" and "U" indicate the subsets of labeled and unlabeled images. PL: Plane, BD: Baseball diamond, BR: Bridge, GTF: Ground track field, SV: Small vehicle, LV: Large vehicle, SH: Ship, TC: Tennis court, BC: Basketball court, ST: Storage tank, SBF: Soccer-ball field, RA: Roundabout, HA: Harbor, SP: Swimming pool, HC: Helicopter.

Category	10% 1179		20% 2195		30% 3251	
	L	U	L	U	L	U
PL	1454	12923	3402	10975	5645	8740
BD	112	614	200	526	250	476
BR	378	3074	839	2613	1324	2128
GTF	54	482	118	418	161	375
SV	2589	196695	47759	174025	79260	148010
LV	2925	40959	6176	37780	12070	31228
SH	549	61549	10307	56791	16336	50762
TC	85	678	135	628	198	565
BC	50	889	109	830	219	720
ST	787	8586	2166	7207	3224	6149
SBF	38	548	119	467	211	375
RA	85	678	135	628	198	565
HA	1102	10255	1921	9436	2738	8619
SP	390	3495	949	2936	1326	2559
HC	57	1029	290	796	294	792

SODA-A [16]. SODA-A is a recently proposed benchmark dataset, specifically designed for detecting small objects in aerial images. Similar to DOTA, it comprises a total of 2513 images, divided into three subsets: train set, validation set, and test set, with approximate proportions of 40%, 25%, and 35% respectively. The format of the bounding box (bbox) annotations is also oriented. In SODA-A, nine object classes are annotated: airplane, helicopter, small-vehicle, large-vehicle, ship, container, storage-tank, swimming-pool, and windmill. The average number of instances per image in SODA-A is more than twice that of DOTA, with 159.18 instances in DOTA and 347.02 in SODA-A. The total number of annotated instances is 872,069, with approximately 96% of instances being smaller than 32 $\times$ 32 pixels. The original images are cropped into 800 $\times$ 800 patches with a stride of 650 pixels.

TABLE II: Number of instances in the SODA-A Dataset under annotated proportions of 10%, 20% and 30%, respectively. "L" and "U" indicate the subsets of labeled and unlabeled images. PL: Plane, HC: Helicopter, SV: Small vehicle, LV: Large vehicle, SH: Ship, CN: Container, ST: Storage tank, SP: Swimming pool, WM: Windmill.

Category	10% 2036		20% 4108		30% 6194	
	L	U	L	U	L	U
PL	2143	18399	3622	16920	5847	14695
HC	64	1020	231	853	335	749
SV	35926	321271	71745	285452	107516	249681
LV	1317	13024	2988	11353	3900	10441
SH	3379	36188	7293	32274	11539	28082
CN	10530	98087	21803	86814	32654	75963
ST	1590	25329	4778	22141	8150	18769
SP	2428	20284	4553	18159	6923	15789
WM	1673	15113	3460	13326	4998	11788

During training and testing, the patch size is maintained at 800 $\times$ 800, without upsampling to 1200 $\times$ 1200 as in previous work [16]. This practice increases the detection difficulty but is deemed beneficial for real-world applications.

Data partition. As seen in Table I and II, following previous works [10], [11], this paper utilizes "partially labeled data" to simulate a scenario with limited data annotations. Specifically, 10%, 20%, and 30% of the images from the train sets of both datasets are randomly selected as labeled data, with the remaining images designated as unlabeled. For all experiments, evaluations are performed on the validation set, and the standard mean average precision (mAP) is reported as the evaluation metric.

Implementation details. To make a fair comparison [11], [13], we employ the Rotate FCOS [38] as the anchor-free detector for both the senior and student models, utilizing a ResNet-50 [31] with a Feature Pyramid Network (FPN) [30] as the feature extractor in all experiments. The supervisor model is pretrained using the Soft Teacher method with 10% annotated data, without the addition of extra images or annotations. We apply strong, weak, and normal augmentations as described in [10], [8], with the exception of random resize and random translate, as these may compromise the visual features of small objects in the datasets. The models are trained over 17 epochs on four RTX 4090 GPUs. Using the SGD optimizer, the initial learning rate of 0.0025 is reduced by a factor of ten at the 13th and 16th epochs. Momentum and weight decay are set at 0.9 and 0.0001, respectively. The batch size per GPU is 5, maintaining a 1:4 ratio of labeled to unlabeled data.

In satellite remote sensing, small oriented objects are the majority and densely distributed, where even a slight angle difference can lead to a low IoU of matching. To address this challenging issue, angle jittering is not applied to the supervisor model, as it focuses only on positional and scale jittering when evaluating the reliability of the predicted bounding box. Although R-FCOS benefits from being anchor-free, there are still few matched instances of prediction, due mainly to the small scale, large aspect ratio, and oriented nature of the objects. Therefore, rather than discarding the prediction of the center as in Unbiased Teacher v2, we retain this branch to reduce the associated confidence of background instances for robustness.

B. Ablation Studies

We initially conduct ablation studies to validate the efficacy of each sub-module before comparing our method with other semi-supervised object detection methods. Experiments are carried out on the DOTA v1.5 dataset using only 10% annotated data. Following prior research [10], [8], [9], models trained with partially annotated images serve as our baselines. To assess the effectiveness of the proposed global burn-in, we compare the mean Average Precision (mAP) of training with and without the lateral-learning strategy. When lateral learning is not employed, the weight of the senior model is updated at the start of training. Furthermore, when the online bounding box (bbox) correction is not implemented, the senior model does not participate in pseudo label prediction, effectively acting as a "buffer" for the student model.

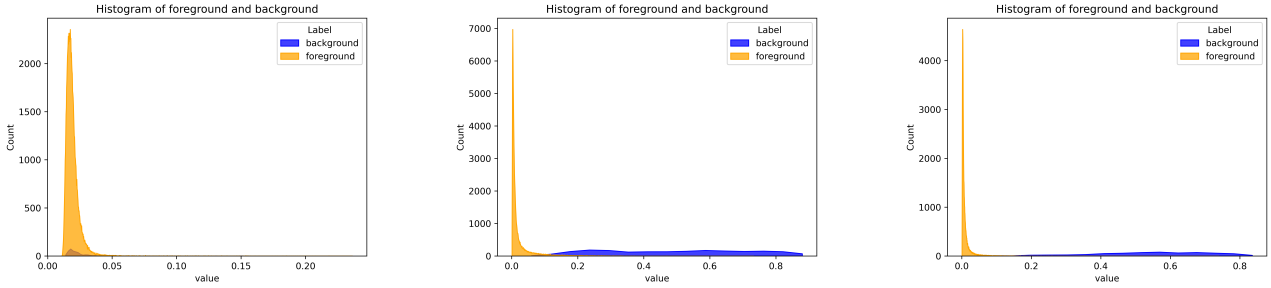
As compared in Table III, with a limited and biased annotated dataset, the oriented RCNN (a two-stage detector) outperforms the Oriented FCOS (a one-stage detector) by approximately 11%. Two-stage detectors, benefiting from positive-negative sampling in the region proposal network, exhibit greater robustness against unbalanced data and inaccurate predictions. We also analyze the prediction range using both a fully annotated and a 10% partially annotated DOTA dataset, as illustrated in Fig. 5. In both scenarios, although the background regions can be accurately classified post-training, the confidence levels for foreground instances remain low, leading to numerous false negative predictions.

Effect of the global burn-in. With the aid of the supervisor model, the proposed method improves performance by about 30%, highlighting the significance of training with an unbiased dataset. Both two-stage detectors and end-to-end learning models typically learn feature extraction from annotated samples, which limits their robustness to variability of object appearance. The supervisor model provides reliable pseudo labels to the student model at the beginning of training, effectively leveraging the potential of unlabeled images. Updating the weight of the senior model at the first step is suboptimal. As shown in Fig. 5, even fully supervised learning models struggle to distinguish objects from backgrounds. Ideally, errors introduced at the beginning would gradually be corrected in later training stages. However, during this phase, prediction errors from the teacher model—specifically the senior model in the Dual Teacher approach—are transferred to the student model, complicating the optimization process and destabilizing the training. By implementing the proposed lateral learning strategy, the mean Average Precision (mAP) is enhanced by 0.7%, indicating a reduction in the errors caused by pseudo labels.

Effect of the online bbox correction. With only 10% of the images annotated, the proposed online bounding box correc-

TABLE III: Ablation studies of the Dual Teacher on the DOTA dataset with 10% annotated data are presented as follows. The Average Precision (AP) for each class is reported, including Plane (PL), Baseball diamond (BD), Bridge (BR), Ground track field (GTF), Small vehicle (SV), Large vehicle (LV), Ship (SH), Tennis court (TC), Basketball court (BC), Storage tank (ST), Soccer-ball field (SBF), Roundabout (RA), Harbor (HA), Swimming pool (SP), and Helicopter (HC). The label "O-" indicates an oriented model. "GBI" and "OBC" refer to the global burn-in and online bounding box correction methods, respectively. "Sup Mod." denotes the use of the supervisor model in training, and "LL" stands for lateral learning.

Methods		PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Baseline	O-FCOS	43.6	4.9	4.6	0.1	17.8	14.6	43.7	47.5	0.4	37.1	0.1	0.2	4.2	21.5	0.0	15.0
	O-RCNN	58.5	21.2	9.8	1.1	26.6	24.1	49.5	64.7	8.6	40.4	1.1	40.4	10.5	18.0	0.0	23.4
+GBI	+Sup Mod.	87.0	67.7	**36.6**	36.2	51.4	72.3	**86.9**	90.1	51.9	59.8	**39.7**	61.0	49.6	**58.5**	12.4	53.8
	+LL	87.6	66.3	33.5	36.8	52.5	70.9	86.7	90.0	**58.5**	60.0	38.1	62.7	50.9	58.4	19.4	54.5
+OBC		**87.8**	**69.9**	36.1	**38.8**	**53.8**	**72.2**	80.7	**90.2**	58.1	**60.1**	37.5	**63.3**	**57.7**	**59.7**	**26.0**	**55.7**



(a) 1st epoch (b) 6th epoch (c) 12th epoch

Fig. 5: Histogram of prediction confidence for the Oriented FCOS during training, using the fully annotated DOTA dataset. Foreground predictions are highlighted in orange and background predictions in blue. Training spans 12 epochs, consistent with prior research. Plots are captured at three key points: (a) the first epoch, (b) the sixth epoch, and (c) the final epoch.

TABLE IV: Comparing the mAP (%) for the Dual Teacher method on the DOTA dataset against other state-of-the-art methods. The abbreviations "O-" and "R-" represent "Oriented" and "Rotate-", respectively, indicating the type of object detection approach used. †indicates the results re-implemented through our training setting

Method	Detector	Param. (M)	FLOPs (G)	mAP (%)		
				10%	20%	30%
Unbiased Teacher [8]	R-RCNN	41.1	211.3	44.8	53.0	52.9
PST†[12]				50.1	57.1	59.4
Soft Teacher [10]				48.0	53.5	54.3
	O-RCNN	41.1	211.4	51.4	56.8	55.7
Dense Teacher [13]	R-FCOS	31.9	206.9	47.1	53.8	56.0
SOOD†[11]				47.8	54.1	56.3
DDPLS†[54]				51.6	56.8	58.1
Focal Teacher†[55]				52.9	57.3	58.7
Supervised (100%)				58.9		
Dual Teacher (Ours)				**55.7**	59.2	59.3
	YOLOx-s	8.94	34.1	51.9	52.7	53.1
	ConvNext	35.8	211.6	55.6	**60.7**	**62.0**

tion method has improved the mean Average Precision (mAP) by 1.2%, and notably enhanced the detection capabilities for small objects such as bridges (+3.6%), small vehicles (+1.3%), and helicopters (+5.6%). The focal loss, which adjusts the contributions of samples based on their "difficulty levels," enables faster network training. However, this benefit assumes that all annotations are "reliable." In semi-supervised learning scenarios, the presence of numerous false negatives can cause focal loss to amplify errors, potentially leading to training collapse. The online bbox correction method, by integrating predictions from two teacher models, mitigates prediction errors that might arise from a single teacher model pipeline. Furthermore, as shown in Fig. 6, during manual annotation, some objects may be misannotated due to factors such as scale, occlusion, or shallow visibility. The proposed online bbox correction strategy can re-annotate these missed instances, thereby reducing the label noise introduced by manual annotations.

Notably, the Average Precision (AP) of the container crane is not reported as it remains zero across all tests, consistent with other semi-supervised learning methods [10], [7], [11], [8]. This phenomenon is attributed to the extremely rare instances of this class in both the partially labeled subset and the validation set, with fewer than 100 instances present. Even under fully supervised learning conditions, the AP for the container crane reaches only 0.1%.

C. Result Comparison

In this subsection, we compare the proposed Dual Teacher with other semi-supervised learning methods. Seven semi-

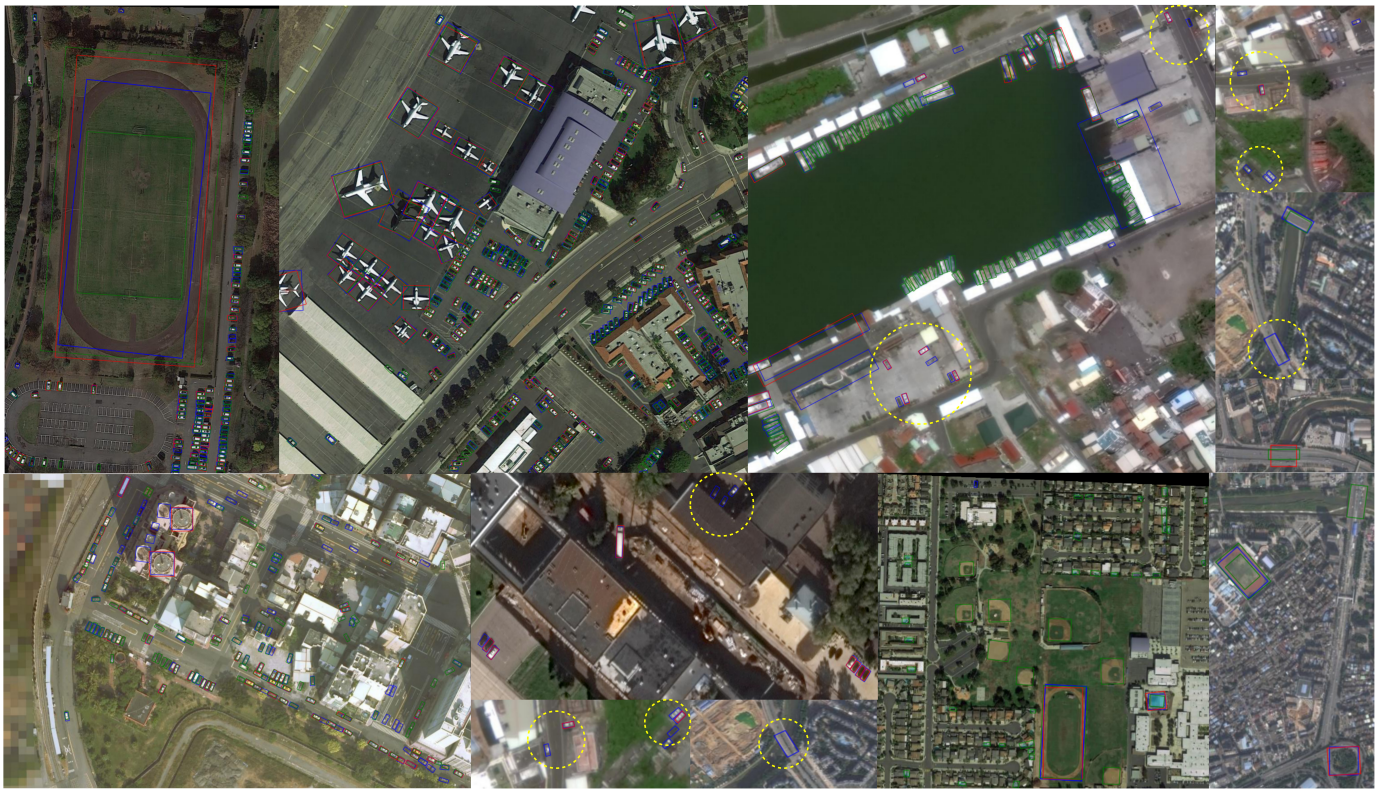


Fig. 6: Visualization of prediction results on the DOTA validation set, where models are optimized using only a 10% annotated subset. Predicted bounding boxes (bboxes) from the supervisor model (Soft Teacher) are highlighted in red, and those from the senior model are in blue. Ground truth bboxes are labeled in green to better illustrate the quality of the bbox predictions. Notably, some instances that are potential foreground objects but not annotated are encircled in dashed yellow. This highlights the Dual Teacher method's robustness to variations in object appearances.

TABLE V: Comparison of results for the Dual Teacher method (%) on the SODA-A dataset against other state-of-the-art methods. The abbreviations "O-" and "R-" represent "Oriented" and "Rotate-", respectively, indicating the type of object detection approach used. †indicates the results re-implemented through our training setting

Method	Detector	mAP		
		10%	20%	30%
Unbiased Teacher [8]	R-RCNN	60.4	64.7	67.6
PST†[12]		64.0	69.9	70.3
Soft Teacher [10]		62.3	65.6	66.8
	O-RCNN	64.6	66.2	67.1
Dense Teacher [13]	R-FCOS	61.9	66.1	67.5
SOOD [11]		62.7	68.6	69.1
DDPLS†[54]		63.5	69.2	70.9
Focal Teacher†[55]		63.8	69.5	71.7
Supervised (100%)	R-FCOS	71.1		
Dual Teacher (Ours)		72.7	73.7	74.1
	YOLOx-s	68.9	69.5	71.2
	ConvNext	**73.5**	**74.8**	**75.3**

supervised oriented object detection methods are selected, which are Unbiased Teacher [8], Soft Teacher [10], PST [12], Dense Teacher [13], SOOD [11], Focal Teacher [55] and DDPLS [54], respectively. The compared methods contains both two-stage [8], [10], [12] and one-stage [11], [13], [54], [55] detectors with varied learning strategies. As shown in Table IV, Dual Teacher surpasses other methods by 4.3% with only 10% labeled data. With 20% labeled data, it outperforms traditional supervised learning by 0.3%. This improvement can be attributed to two factors. Firstly, as previously mentioned, the presence of unlabeled objects can hinder network optimization under a supervised learning framework, whereas Dual Teacher can relabel these for training. Secondly, pseudo label-based methods promote learning consistency between two augmentations, which enhances the detector's robustness in varying scenes. However, with 30% labeled data, the improvement becomes marginal, possibly due to the limited distribution of the additionally annotated data and the method's inability to correct false predictions effectively. A similar pattern is observed with Soft Teacher, where mAP decreases with 30% labeled data. This suggests that some annotations in DOTA might be "redundant annotations," i.e., they do not significantly contribute to training and warrant further investigation. Additionally, we plot the confusion matrices for state-of-the-art methods. As illustrated in Fig. 7, Dual Teacher significantly reduces the number of misannotated instances, particularly for small objects.

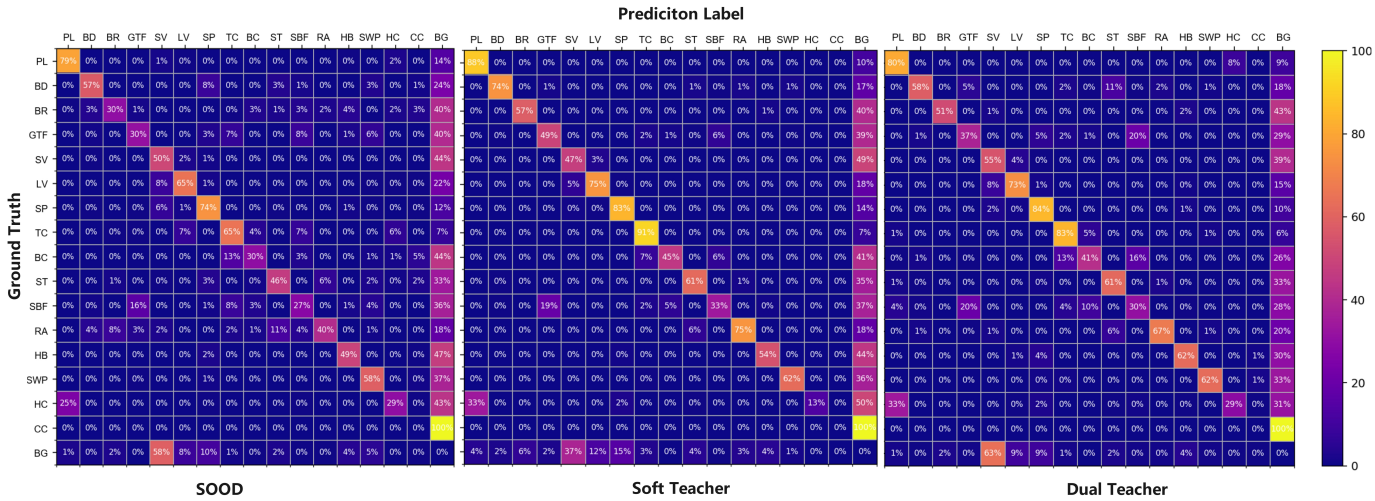


Fig. 7: The confusion matrices for SOOD, Soft Teacher, and Dual Teacher on the DOTA validation set, respectively. All methods are trained with only 10% labeled data, where our Dual Teacher can significantly reduce the false negative rate, especially on the categories of ship (2%), small vehicle (5%), swimming pool (2%), harbor (14%), and helicopter (10%).

TABLE VI: Result comparisons (%) in terms of label types on the DOTA dataset. "OBL" denotes the proposed online bbox correction method, which is not utilized in this test. The proposed global burn-in is applied in Dense Teacher and SOOD. Oriented FCOS with ResNet-50 is employed across all methods to ensure a fair comparison.

Method	Label type	mAP		
		10%	20%	30%
Dense Teacher	Logit	47.9 (↑ 0.8)	54.2 (↑ 0.4)	56.7 (↑ 0.7)
SOOD		48.0 (↑ 0.2)	54.6 (↑ 0.5)	56.9 (↑ 0.6)
Dual Teacher (Ours) w/o OBL	Pseudo label	**54.5**	**57.3**	**57.9**

When evaluated on the SODA-A dataset, as shown in Table V, our Dual Teacher method also surpasses other methods, exceeding the performance of supervised training with 20% labeled data by 2.6%. The SODA-A dataset specifically focuses on detecting small objects. Results are reported without scale sampling, which underscores the method's effectiveness in detecting small objects within the realm of remote sensing.

D. Extended Discussions

1) Soft Label vs. Hard Label: As discussed in Section II, some semi-supervised learning methods utilize "soft-labels," where pseudo labels are derived from the feature maps of the teacher model. We evaluated this approach within our proposed learning pipeline. Following the global burn-in phase, the supervisor model is removed, and the standard focal loss is replaced with the Quality Focal Loss [56] or the Rotation-aware Adaptive Weighting loss combined with Global Consistency loss, as used in Dense Teacher [13] and SOOD, respectively. This adaptation allows for the learning of continuous dense pseudo-labels. As shown in Table VI, there are slight mean Average Precision (mAP) improvements compared to their respective conventional methods. We attribute these improvements to the distinct divergence in object properties between general and remote sensing scenes. In remote sensing, the image scales are larger and object scales are smaller, which exacerbates the imbalance between positive and negative labels. Additionally, factors such as occlusion and blurriness reduce the prediction confidence of the teacher model, which is detrimental to semi-supervised learning. Consequently, employing soft-labels and learning with low-confidence pseudo labels can increase the prediction uncertainty of the student model.

In contrast, the "hard label" learning approach, utilized in the proposed Dual Teacher method, excludes low-quality pseudo labels and reassigns the label to "1", effectively reducing prediction ambiguity and clarifying the learning target. This method is particularly suitable for remote sensing scenes, where class imbalance and environmental interference are more pronounced [29], [57]. This strategy helps to sharpen the focus of the model on more reliable data, thereby enhancing overall performance.

2) Self-supervised Learning vs. Global burn-in: At the beginning of training, the global burn-in serves to prevent the training process from converging to a low-quality local minimum, effectively enriching the prior knowledge of the senior model. This concept is analogous to fine-tuning a model that has been trained via self-supervised learning. To evaluate the efficacy of self-supervised learning versus the proposed global burn-in, we conduct experiments on knowledge inheritance among different methods, where the detector is initialized either with global burn-in or with self-supervised pretrained weights. As shown in Table VII, the maximum drop in mean Average Precision (mAP) is about 5.2%, when compared to the proposed global burn-in. Even when compared to their original burn-in settings, where the weights of the backbone are trained with supervised learning on ImageNet, the mAP

TABLE VII: Result comparisons (%) in terms of knowledge inheritance methods on a 10% labeled DOTA dataset. "GBI" denotes the proposed global burn-in method. "Ori." refers to the original burn-in methods, which include two-stage learning for Dense Teacher and SOOD, and end-to-end learning for Soft Teacher.

Method	Burn-in Strategy	mAP	Δ
Dense Teacher	Ori.	47.1	-
	MoCo	43.2	-3.9
	BYOL	45.5	-1.1
	GBI	47.9	+0.8
Soft Teacher	Ori.	51.4	-
	MoCo	48.1	-3.3
	BYOL	50.0	-1.4
	GBI	53.3	+1.9
SOOD	Ori.	47.8	-
	MoCo	44.0	-3.0
	BYOL	46.4	-1.4
	GBI	48.0	+0.2

still decreases by 1% to 4%. We deduce that this is mainly caused by the alternation of the learning target.

The primary objective of object detection is to localize and classify objects within images. Fine-tuning a model that has been trained using supervised methods also capitalizes on its inherent capabilities in classification. In contrast, self-supervised training methods often focus on a learning target that is fundamentally different from the tasks of detection. This discrepancy can significantly hinder weight optimization, a challenge also noted in the literature [58], [59]. Additionally, shifts in dataset characteristics and the noise inherent in pseudo-labels further compound the difficulty of learning.

As a comparison, the proposed global burn-in facilitates the inheritance of knowledge from another detector with the same learning target, potentially reducing training difficulties. When integrated into "hard label" learning frameworks, global burn-in can further enhance performance by approximately 2%, thereby validating its effectiveness and robustness.

3) Comparisons of BBOX Merging strategies: Although vanilla Non-Maximum Suppression (NMS) is employed for merging bounding boxes (bboxes) between the supervisor model and the senior model, we have also explored alternative merging strategies, namely soft NMS and the "union merge." In the union merge method, a bbox that matches with the highest Intersection over Union (IoU) value with the reference bbox is merged into a "union" of the two bboxes. The reference merged bbox then remains unchanged in subsequent IoU comparisons, and other matched bboxes are discarded as in vanilla NMS. Experimental results, as presented in Table VIII, show that both methods lag behind vanilla NMS by less than 1%. For soft NMS, while the confidences of duplicate bboxes are reduced, the re-weighted bbox might be filtered out due to the original low confidences predicted from the senior model. The "union merge" method only adjusts bbox size, and since most predictions from the two teacher models are similar, the overall bbox predictions are not significantly varied. Additionally, both methods increase computational costs, consequently slowing down the training process.

TABLE VIII: Result comparisons (%) of different bbox merging methods on DOTA dataset.

Method	mAP		
	10%	20%	30%
Soft NMS	54.9	58.8	58.6
Merge	55.2	59.1	58.7
Vanilla NMS	55.7	59.2	59.3

TABLE IX: Comparing the mAP (%), number of parameters (M) and FLOPs (G) from different training strategies using 30% of annotated DOTA dataset. "1x" and "2x" denotes models are trained using the original setting and with doubled iterations, respectively. "O-" and "R-" represent "Oriented" and "Rotate-", respectively, indicating the type of object detection approach used. "T." is the short of "Teacher". FLOPs are calculated with an input image size of 1024 × 1024. The best result on each mAP column is highlighted in bold, and the second best is underlined. "+" indicates that the "strong augmentation" is applied for data preprocessing.

Mode	Methods	Param. (M)	FLOPs (G)	mAP	
				1x	2x
Sup.	R-FCOS	31.9	206.9	58.9	57.1
	R-FCOS+			54.6	52.9
	R-RCNN	41.1	211.3	59.2	61.5
	R-RCNN+			59.4	61.0
	O-RCNN	41.1	211.4	**64.6**	**63.8**
	O-RCNN+			59.2	59.1
	R-YOLOx-s	8.94	34.1	61.6	59.9
Semi-Sup.	SOOD	31.9	206.9	56.3	58.2
	DDPLS			58.1	60.0
	Dual T.	31.9	206.9	59.3	60.9
	Dual T. (ConvNext)	35.8	211.6	<u>61.9</u>	<u>62.0</u>
	Dual T. (YOLOx-s)	8.94	34.1	53.1	53.0

4) Comparisons of model architecture and training strategies: We also validate the effect of our Dual Teacher method with different model architecture, including RepPoints [39], YOLOX-s [51] and ConvNext (as the backbone) [60]. As seen in TableIX, when utilizing the ConvNext as the backbone, our Dual Teacher lags the supervised O-RCNN by 2.1% on average. It also outperforms the supervised YOLOX-s and R-FCOS by 0.3% and 3%, respectively. When compared with other semi-supervised learning methods, the Dual Teacher out-

TABLE X: Ablation study on the number of senior models. Models are trained with 10% annotated DOTA dataset.

Num. Senior	1	2	3
mAP	**55.7**	55.6	55.5

performs by 2% at minimum. When comparing with YOLO-s as the detector, the mAP lags further, which can be deduced by two folds. Firstly, YOLO models are boosted by the matching strategy, such as SimOTA [51], for encouraging the "anchor point" to match with more foreground objects. This will bring benefits when the labels are reliable. Considering that the pseudo labels are not 100% correctly annotated, it will speed-up the learning of the annotation noise, causing the YOLOx-s hard to be optimized in the semi-supervised learning. This is also validated in Table IV and V. As SODA-A contains more annotated instances than DOTA v1.5, the mAP of YOLOx-s can be improved significantly when the annotated ratio increases. For comparison, the RFCOS learning scheme can stability learn feature extraction well on those two datasets, using either ResNet-50 or ConvNext as the backbone. Similarly, the matching strategy in RepPoints also impedes the model optimization, especially when no objects are detected by the supervisor model. Eventually, the RepPoints cannot be well optimized, which is not shown here. For comparison, the FCOS matches objects with predictions around the central region, without additionally prediction matching, is more robust to the annotation noise.

As shown in Table IV and V, the fully supervised learning methods are unperformed compared to the proposed Dual Teacher model. When compared to other semi-supervised learning methods, we notice that they are trained by about 120k-180K iterations. As the ratio of the labeled and unlabeled images is maintained constant in each batch, the number of training iteration per epoch is $\sim$5000 in our training schedule. As a result, the proposed Dual Teacher is trained with 60K iteration. To validate the effect of the training schedule, we further extend the iterations by double the number of samples within an epoch. As shown in Table IX, only the R-FCOS-based Dual Teacher improves the mAP by 1%. In contrast, the mAPs of DDPLS and SOOD are improved by around 2%, which indicates that the proposed Dual Teacher can be well optimized with fewer training iterations.

The pseudo-label-based semi-supervised learning methods can well learn the consistency between two augmentations. In addition, we also separately validate their effectiveness on supervised learning. As seen in Table IX, with strong augmentation applied, the performance is not improved as expected. On the contrary, it may degrade the mAP in some cases.

5) The number of senior models: Since existing semi-supervised learning methods use one senior model for prediction, we also explore the effectiveness on multiple senior models, where the weights are iteratively updated from the student model. As seen in Table X, the prediction results are not further improved when more than one senior model is added. We deduce that may be caused by the error accumulation. In this study, all senior models are updated through the student models, where the weight differences are minor. As a contrast, the architecture difference between the supervisor model and the senior model are large, this will allow the student model learn the merit from both models.

6) Limitations: Although the proposed Dual Teacher method has achieved promising results, it also has certain limitations. As discussed in Section IV-B, the Dual Teacher can be affected by categories with rare instances. While the two-teacher network can significantly reduce the number of misannotated objects, the occurrence of missed detections remains considerable. Furthermore, compared to two-stage detectors, the prediction confidences of one-stage detectors are typically lower, which can lead to a higher rate of false positives in practical applications. Additionally, the quality of the bounding boxes in certain categories, such as bridges, soccer fields, and harbors, requires further refinement. For the bridge and the soccer field categories, this is possibly caused by the low contrast between the object and the background in some scenarios. For the harbor category, we deduce this is caused by the extra-large aspect ratio. A marginal shift will cause a large drop on the IoU between the prediction and the ground truth.

V. Conclusion and Future Works

In this paper, to enhance the reliability, as well as the accuracy, of pseudo bounding boxes (bboxes) in semi-supervised oriented object detection within the realm of remote sensing, we introduce a novel two-heterogeneous-teacher-based method, named Dual Teacher. Unlike traditional pseudo-label based methods, Dual Teacher incorporates an additional detector, termed the "supervisor" model, to refine pseudo label quality and assist the primary teacher network (referred to as the senior model) in reducing the prediction bias caused by limited annotated data. The supervisor model is optimized through semi-supervised learning, requiring no additional annotations or images, thereby enhancing its usability. Through our experiment, we find out that initialize the detector before applying the teacher-student scheme will boost the prediction quality of pseudo bboxes. Thus, we propose a global burn-in method, where the unlabeled images are annotated by the supervisor model, enabling the detector learning feature presentation through the global scale, i.e., detector is trained among the whole training set despite annotated or not. Detecting small objects in the remote sensing field may encounter the ambiguity of visual features, causing the low prediction confidence. Numerous foreground instances will then be rejected in the previous bbox sampling method. To address this, we propose an online bbox correction scheme that selects potential foreground bboxes from low-prediction-score samples, which remarkably reduce the false negative rate of the pseudo bboxe set. Our proposed Dual Teacher method outperforms traditional supervised learning on both DOTA and SODA-A datasets with only 20% labeled data. Moreover, Dual Teacher significantly reduces the manual labor required for annotation and proves robust against environmental interference and errors in manual annotation.

Although soft-label learning is not currently suitable for the proposed Dual Teacher, its potential in self-supervised learning [61] and general scene detection motivates us to refine its learning scheme for remote sensing images. Future work could involve integrating soft-label learning with a confidence calibration module, or replacing the focal loss with cross-entropy loss [8], [11]. Additionally, existing self-supervised learning methods cannot be directly applied to semi-supervised learning in remote sensing due to significant data shifts between general scene datasets [62], [63] and remote sensing datasets. Thus, future efforts should focus on narrowing this gap by setting a local-feature learning target and mitigating data shifts. Furthermore, an object identification method should be explored to further reduce missed detections, coupled with a pseudo-label refining method to enhance the bbox quality.

REFERENCES

[1] L. Wen, Y. Cheng, Y. Fang, and X. Li, "A comprehensive survey of oriented object detection in remote sensing images," *Expert Systems with Applications*, p. 119960, 2023.

[2] H. F. Tolie, J. Ren, and E. Elyan, "Dicam: Deep inception and channel-wise attention modules for underwater image enhancement," *Neurocomputing*, vol. 584, p. 127585, 2024.

[3] P. Ma, J. Ren, G. Sun, H. Zhao, X. Jia, Y. Yan, and J. Zabalza, "Multiscale superpixelwise prophet model for noise-robust feature extraction in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[4] Y. Li, J. Ren, Y. Yan, Q. Liu, P. Ma, A. Petrovski, and H. Sun, "Cbanet: an end-to-end cross band 2-d attention network for hyperspectral change detection in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[5] Y. Yan, J. Ren, H. Sun, and R. Williams, "Nondestructive quantitative measurement for precision quality control in additive manufacturing using hyperspectral imagery and machine learning," *IEEE Transactions on Industrial Informatics*, 2024.

[6] L. Lei, Z. Fang, J. Ren, P. Gamba, J. Zheng, and H. Zhao, "Two-click based fast small object annotation in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[7] 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.

[8] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," *arXiv preprint arXiv:2102.09480*, 2021.

[9] Y.-C. Liu, C.-Y. Ma, and Z. Kira, "Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9819–9828.

[10] 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.

[11] W. Hua, D. Liang, J. Li, X. Liu, Z. Zou, X. Ye, and X. Bai, "Sood: Towards semi-supervised oriented object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 558–15 567.

[12] W. Wu, H.-S. Wong, and S. Wu, "Pseudo-siamese teacher for semi-supervised oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[13] 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 *European Conference on Computer Vision*. Springer, 2022, pp. 35–50.

[14] Z. Zhang, Y. Wang, C. He, Q. Zhang, and X. Chen, "Weakly semi-supervised oriented object detection with points," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 3080–3084.

[15] 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.

[16] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, and J. Han, "Towards large-scale small object detection: Survey and benchmarks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[17] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *Advances in neural information processing systems*, vol. 27, 2014.

[18] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[19] T. Miyato, S.-i. 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*, vol. 41, no. 8, pp. 1979–1993, 2018.

[20] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representations*, 2016.

[21] 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.

[22] P. Tang, C. Ramaiah, Y. Wang, R. Xu, and C. Xiong, "Proposal learning for semi-supervised object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2291–2301.

[23] 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.

[24] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, "Rethinking pre-training and self-training," *Advances in neural information processing systems*, vol. 33, pp. 3833–3845, 2020.

[25] Y. Li, D. Huang, D. Qin, L. Wang, and B. Gong, "Improving object detection with selective self-supervised self-training," in *European Conference on Computer Vision*. Springer, 2020, pp. 589–607.

[26] 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.

[27] 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.

[28] 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.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[30] 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.

[31] 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.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[33] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3520–3529.

[34] 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.

[35] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with gaussian wasserstein distance loss," in *International conference on machine learning*. PMLR, 2021, pp. 11 830–11 841.

[36] X. Yang, X. Yang, J. Yang, Q. Ming, W. Wang, Q. Tian, and J. Yan, "Learning high-precision bounding box for rotated object detection via kullback-leibler divergence," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 381–18 394, 2021.

[37] X. Yang, Y. Zhou, G. Zhang, J. Yang, W. Wang, J. Yan, X. Zhang, and Q. Tian, "The kfiou loss for rotated object detection," *arXiv preprint arXiv:2201.12558*, 2022.

[38] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: A simple and strong anchor-free object detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1922–1933, 2020.

[39] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9657–9666.

[40] L. Hou, K. Lu, X. Yang, Y. Li, and J. Xue, "G-rep: Gaussian representation for arbitrary-oriented object detection," *Remote Sensing*, 2023.

[41] X. Zheng, H. Cui, C. Xu, and X. Lu, "Dual teacher: A semisupervised cotraining framework for cross-domain ship detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[42] Y. Xin, Z. Fan, X. Qi, Y. Zhang, and X. Li, "Confidence-weighted dual-teacher networks with biased contrastive learning for semi-supervised semantic segmentation in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[43] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[44] G. Cheng, X. Xie, W. Chen, X. Feng, X. Yao, and J. Han, "Self-guided proposal generation for weakly supervised object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[45] Y. Pang, Y. Zhang, Q. Kong, Y. Wang, B. Chen, and X. Cao, "Socdet: A lightweight and accurate oriented object detection network for satellite on-orbit computing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[46] Z. Li, B. Hou, Z. Wu, B. Ren, Z. Ren, and L. Jiao, "Gaussian synthesis for high-precision location in oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[47] S. Zheng, Z. Wu, Y. Xu, and Z. Wei, "Instance-aware spatial-frequency feature fusion detector for oriented object detection in remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[48] Y. Yu, X. Yang, Q. Li, Y. Zhou, F. Da, and J. Yan, "H2rbox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[49] S. Zhang, Z. Yu, L. Liu, X. Wang, A. Zhou, and K. Chen, "Group r-cnn for weakly semi-supervised object detection with points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9417–9426.

[50] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[51] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[52] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.

[53] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 2888–2897.

[54] T. Zhao, Q. Fang, S. Shi, and X. Xu, "Adaptive dense pseudo label selection for semi-supervised oriented object detection," *arXiv preprint arXiv:2311.12608*, 2023.

[55] K. Wang, Z. Xiao, Q. Wan, F. Xia, P. Chen, and D. Li, "Global focal learning for semi-supervised oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[56] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 002–21 012, 2020.

[57] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 761–769.

[58] T. Dang, S. Kornblith, H. T. Nguyen, P. Chin, and M. Khademi, "A study on self-supervised object detection pretraining," in *European Conference on Computer Vision*. Springer, 2022, pp. 86–99.

[59] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, "Rethinking pre-training and self-training," *Advances in neural information processing systems*, vol. 33, pp. 3833–3845, 2020.

[60] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 133–16 142.

[61] X. Wang, R. Zhang, C. Shen, and T. Kong, "Densecl: A simple framework for self-supervised dense visual pre-training," *Visual Informatics*, vol. 7, no. 1, pp. 30–40, 2023.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[63] 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.

Zhenyu Fang received the Ph.D. degree in electronic and electrical engineering from the University of Strathclyde in July 2020. He is a currently an Associate Professor with the the School of Software, Northwestern Polytechnical University, Xi'an, China, and also an Associate Professor with Yangtze River Delta Research Institute of NPU, Taicang, China. His main interests are algorithm development for small object detection, self-supervised learning, semi-supervised learning and model compression.

Jinchang Ren (Senior Member, IEEE) the Ph.D. degree in electronic imaging and media communication from the University of Bradford, Bradford, U.K., in 2009. He is a Professor with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K., and also a Visiting Professor with Guangdong Polytechnic Normal University, Guangzhou. He has published over 350 articles. His research interests include computer vision and multimedia signal processing, especially on hyperspectral imaging, machine learning, and big data analytics.

Jiangbin Zheng received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, Shaanxi, China, in 2002.,He is currently a Full Professor and the Dean of the School of Software, Northwestern Polytechnical University. His research interests include computer graphics, computer vision, and multimedia. He has authored over 100 articles in the above-related research area.

Rongjun Chen Rongjun Chen received the M.S. degree in control theory and control engineering from Guangdong University of Technology, Guangzhou, China, in 2007, and the Ph.D. degree in communication and information system from Sun Yat-sen University, Guangzhou, in 2015.,He is currently a Professor with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou. His research interests include image perception and processing, as well as the Internet of Things.

Huimin Zhao received the B.Sc. and M.Sc. degrees in signal processing from Northwestern Polytechnical University, Xi'an, China, in 1992 and 1997, respectively, and the Ph.D. degree in electrical engineering from Sun Yat-sen University, Guangzhou, China, in 2001.,He is currently a Professor and Dean with the School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou. His research interests include image/video and information security technologies, and applications.