Boosting Semi-Supervised Object Detection in Remote Sensing Images with Active Teaching

IEEE Publication Technology, *Staff, IEEE,*

***Abstract*—The lack of object-level annotations is a main challenge for object detection in remote sensing images. Active learning and semi-supervised learning can improve the quality and quantity of annotations by identifying the most informative samples for annotation and exploring the knowledge from the unlabeled samples respectively. In this paper, we propose a novel semi-supervised object detection method with active teaching for remote sensing images named SSOD-AT by combining object-level pseudo labeling and informative active annotation with a teacher-student network. In the proposed method, a RoI Comparison module (RoICM) is designed based on the teacher-student framework to provide high-confident pseudo-labels of RoIs. Meanwhile, we also use the RoICM to identify the top-K uncertain images. Then a diversity criterion is adopted based on the object-level prototypes of different categories with the labeled images and the pseudo-labeled images to remove the redundancy in the top-K uncertain images for human labeling. The extensive experiments on two popular datasets DOTA and DIOR show that the proposed method outperforms the state-of-the-art methods.**

***Index Terms*—Active learning(AL), semi-supervised object detection(SSOD), teacher-student framework, remote sensing.**

I. Introduction

OBJECT detection in remote sensing images is an essential task in computer vision to discover that what and where are the objects in the images [1]. In recent years, deep learning has achieved satisfactory performance for object detection in remote sensing images. Generally, the success of the approaches with deep learning usually depend on the large-scale datasets with high-quality labels from the human experts annotation. However, compared with the classification task, which requires to annotate images with the image-level labels, the annotation for object detection is much cost due to that it needs object-level labels with both bounding box and categories. Meanwhile, in remote sensing image, the objects are usually with different orientations and scales [2], [3], [4], [5], [6], which further increase the efforts for labeling. Hence, with the rapid development of remote sensing technique, although the collection of the remote sensing images becomes more and more fast and easy, the available labeled images for object detection are very limited [7], [8], [9].

Semi-supervised learning (SSL) and active learning (AL) are two promising techniques in machine learning to address the problem with limited labeled images. SSL usually attempts to exploit the unlabeled data with a limited amount of labeled data by assuming the consistency between the feature distribution of unlabeled data and labeled data. Pseudo-labeling is a popular strategy for SSL by selecting the samples with high-confident predictions [10]. SSL with Pseudo-labeling has achieved satisfactory performance in classification [11], [12], [13], [14] and object detection [15], [16], [17]. For the semi-supervised object detection method, most of them are developed based on a teacher-student network, which introduces a secondary model (teacher) to guide the training of the primary model (student). The teacher network uses weakly augmented labeled data to generate high-quality pseudo-labels for the student network [18], [19], [20]. The student network then is optimized by using the pseudo-labels as supervised information. This self-training procedure enables the model to fully utilize the large quantity of unlabeled data with a very small number of labeled samples.

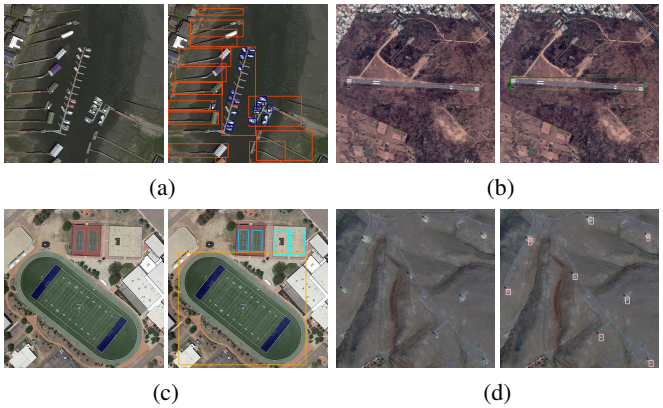


Fig. 1. Examples in RSIs with objects that increase labeling effort. (a) Objects with multiple orientation. (b) Objects with extreme aspect ratio. (c) Objects with large-scale variations. (d) Objects with tiny sizes.

Different from SSL, AL generally focuses on the labeled data, which are mainly collected by selecting the most informative samples from the unlabeled data for human experts labeling [21]. For the setting of AL, it initializes a small number of labeled samples to train a supervised model. Then the query criterion is designed to select the most informative samples with an iterative manner. Hence, the query criterion is the core technique in AL methods, i.e. uncertainty and diversity [22], [23]. The uncertainty mainly focuses on the samples that are hard to make a confident decision with the model, while the diversity is adopted to remove the redundancy within the labeled samples. For instance, Yuan et.al [24] proposes MI-AOD, an instance-level uncertainty-based method which highlights the informative instances while filtering out noisy ones to select the most informative images for detector training. CALD [25] not only gauges individual information for sample selection but also leverages mutual information to encourage a balanced data distribution, thus alleviating unbalanced class distribution to ensure the diversity of selected

samples.

We note that SSL relies on labeled data to predict high-confidence pseudo-labels, while AL is able to select the most informative samples for labeling, thus removing the redundancy in the labeled set, which indicates their complementary and synergistic. Hence, it is a natural consideration to combine active learning and semi-supervised learning together to improve the performance of object detection. In fact, many methods has developed for SSOD with active learning. Lv et.al proposed a novel semi-supervised active salient object detection(SOD) method, which first designed a saliency encoder-decoder with adversarial discriminator to select the least confident (discriminative) samples to form the "candidate labeled pool". Then, SOD train a VAE to select and add the most representative data from the "candidate labeled pool" into the labeled pool by comparing their corresponding features in the latent space. Mi et al [26] presented the first attempt of studying data initialization by AL in teacher-student based SSOD, called *Active Teacher*. In *Active Teacher*, the label set is partially initialized and gradually augmented via a novel active sampling strategy from the aspects of difficulty, information and diversity. Despite the substantial progress, these combinatorial methods still have some non-negligible deficiencies: (1) Their query criterions generally depend on accurately predicted RoIs (e.g., pseudo-labels predicted by the teacher network with high confidence). Due to the relatively complex object distribution in RSIs, pseudo-labeling may only produce a few credible pseudo-labels, especially at the early stage of the training process, which will bias the calculation of uncertainty and ignore truly informative (e.g., hard-to-detect) samples. (2) Their query criterions usually require querying the entire unlabeled set to select the most informative samples for expert labeling, which is a time-consuming workload for remote sensing datasets with a large quantity of images.

In this paper, aimed at the characteristics of remote-sensing images, we proposed a method which boost Semi-supervised object detection by active teaching(SSOD-AT). The proposed method can effectively integrate semi-supervised and active learning, allowing the model to make full use of information from unlabeled data while also obtaining reliable and high-quality labeled training samples for iterative training. Since objects in remote sensing images tend to be characterized by large scale variations and high density, the pseudo-labels generated by the teacher network in the traditional teacher-student framework may not be able to accurately locate the objects in the images, and the unreliability of the pseudo-labels may in turn significantly affect the training of the student network. As a result, we proposed a RoI comparison module(RoICM) in which we compare the RoIs predicted by the teacher and student networks of the same unlabeled image before the teacher network's pseudo-labeling process, and only images with consistent RoIs are retained for the teacher network to generate pseudo-labels, effectively ensuring the reliability of the pseudo-labels. The pseudo-labels generated by teacher network, on the other hand, are deemed untrustworthy if the RoIs are inconsistent, which contributes a degree of uncertainty to the unlabeled image. Typically, uncertainty serves as one of the key factors for selecting samples for labeling in active learning, and the more uncertain the image, the more valuable it is to be actively labeled. Based on this, we calculate the uncertainty of unlabeled images with inconsistent RoIs by two networks as one of the metrics for active learning sampling. Another crucial metric for active learning is class diversity of actively selected image, which is something we ensure based on the concept of class prototypes. To maximise the effectiveness of active learning, we integrate both metrics as the final screening weights for unlabeled images, which are used to choose the most valuable images for human labeling. The main contributions of this article can be summarized as follows.

- A novel active learning teacher-student framework, named SSOD-AT, is proposed for remote-sensing object detection. With this advanced paradigm that overcomes the crucial challenges, SSOD-AT can achieve high detection accuracy and robustness only with limited labeled samples.
- The proposed SSOD-AT introduces a RoI comparison module(RoICM) which compares the predictions generated by the two networks. On the one hand, it guarantees the accuracy of the pseudo-labels, and on the other, it can be used to evaluate the image uncertainty for AL sampling.
- The proposed SSOD-AT further incorporates the global class prototype for the image diversity in AL to ensure the class diversity of the selected samples. The combination of the two sampling strategies maximizes the effectiveness of AL process.

The rest of this paper is organized as follows. Section II presents details of the related works on remote sensing image object detection, semi-supervised learning and active learning. Section III describes the proposed SSOD-AT algorithm in detail. Section IV describes the verification of the effectiveness of SSOD-AT. Section V summarizes this paper.

II. Related Works

Object detection in remote sensing images (RSIs) requires the locating and classifying of objects of interest, which is a hot topic in RSI analysis research due to its important role in various applications. In this section, first, we briefly review recent object detection techniques in remote-sensing images and, then, discuss two important branches of deep learning, semi-supervised learning and active learning applied in object detection.

A. Object detection in remote-sensing images

With the development of deep learning technology, which has accelerated in recent years, numerous intelligent and efficient object detection algorithms have been proposed. Mainstream object detection algorithms can be roughly divided into two categories: one-stage and two-stage, with the main difference being whether they include a step for the proposed region. The two-stage algorithms are represented by R-CNN family, a collection of algorithms. Among them, Faster R-CNN [27] is one of the most popular method, which first

designed the RPN subnet to obtain proposals from the anchor mechanism. Faster R-CNN is also frequently employed as the bottom network in remote-sensing object detection. Representative one-stage algorithms include the anchor-free families [28], [29]. CornerNet [28] proposed by Law et al. and CenterNet [29] proposed by Duan et al. generate bounding boxes by detecting the corner points or center points of objects to further accelerate speed detection. They have become a new research topic in remote sensing.

However, within remote sensing images, geographical objects might manifest distinct visual characteristics across diverse orientations, angles, and scales. Consequently, the common object detection algorithms encounter challenges in their applicability. Therefore, a range of improvement strategies and optimization methods are proposed for remote sensing image interpretation [30], [31], [32], [33]. Chen et al. [30] proposed a cascade attention network (CA-CNN) composed of a patched self-attention module and a supervised spatial attention module to improve the feature representation of objects. Cui et al. [31] designed a new anchor-free remote sensing ship detection model named SKNet, which constructs an orthogonal pool to highlight the features of the central point and its surroundings. On this basis, it then predicts the morphology of the central point. Currently, with the development of Transformer, it is extended to various computer vision fields, including remote sensing object detection. Zhang et al. [32] introduced a novel coarse-to-fine framework (CoF-Net). CoF-Net mainly consists of two parallel branches, namely coarse-to-fine feature adaptation (CoF-FA) and coarse-to-fine sample assignment (CoF-SA), which aim to progressively enhance feature representation and select stronger training samples, respectively. Additionally, oriented detection, which can offer objects' orientation and scale information, is widely employed in remote sensing image interpretation. The most recent contribution originates from the investigation conducted by Zhang et al. [33], who proposed a task collaborated oriented detector(TCD) to make predictions with high quality in both classification and localization.

B. Semi-Supervised object detection

In the existing research, semi-supervised learning are mainly applied to image classification tasks [13], [14], [34], [35], [36], [37], [38], [39], [40], [41], which are broadly classified into the following categories: consistency regularization, proxy-label and Generative models, respectively. Consistency regularization methods [13], [14], [35], [36] based on the assumption that if a realistic perturbation was applied to the unlabeled data points, the prediction should not change significantly, thus the model can be trained to have a consistent prediction on a given unlabeled example and its perturbed version. Proxy-label methods [14], [34], [37], [38], [39] leverage a trained model on the labeled set to produce additional training examples by labeling instances of the unlabeled set based on some heuristics. Similar to the supervised setting, where the learned features on one task can be transferred to other downstream tasks, Generative model methods [40], [41] that are able to generate images from the data distribution must learn transferable features to a supervised task for a given task with targets.

Recent research has also focused extensively on the application of semi-supervised learning to object detection(SSOD), and numerous superb methods have been developed. In particular, the teacher-student is one of the most popular underlying framework. The first teacher-student based framework for SSOD is proposed by STAC [16], which deployed highly confident pseudo labels of localized objects from an unlabeled image and updates the model by enforcing consistency via strong augmentations. Zhou et al. designed Instant-Teaching [42] which uses instant pseudo labeling with extended weak-strong data augmentations for teaching during each training iteration. Nevertheless, the aforementioned methods still suffer from excessive instability during the initial training phase and necessitate a high confidence score threshold for generating pseudo-labels. Unbiased teacher [15] proposed by Liu et al. jointly trained a student and a gradually progressing teacher in a mutually-beneficial manner called EMA [13], together with a class-balance loss called focal loss [43] to downweight overly confident pseudo-labels. Considering the pseudo-labeling by thresholding is not the best solution since the confidence value is not strictly related to the prediction uncertainty, ILnet [44] proposed by Rossi L et al. introduced an additional classification task for bounding box localization to improve the filtering of the predicted bounding boxes and obtain higher quality on Student training, which empirically proves that bounding box regression on the unsupervised part can equally contribute to the training as much as category classification.

C. Active Learning in object detection

In order to maximise accuracy while using the fewest number of requests possible, active learning [45] seeks to carefully select the samples to be labeled, lowering the cost of acquiring labeled data. The two most common selection criteria are representativeness and informativeness [46]. Representativeness assesses how effectively an instance helps represent the structure of input patterns, whereas informativeness measures how well an unlabeled instance helps minimise the uncertainty of a statistical model. In order to decrease the cost of labeling for object detection, certain active-learning based approaches [25], [24] have also been proposed. CALD proposed by Yu et al. [25] fully explores the consistency between the original and augmented data, which designed as the information of samples. Yuan et al. [24] proposed active learning method based on multi-instance learning called MI-AOD, which selects the most informative images for detector training by observing instance-level uncertainty of samples.

Since both aim to use a limited amount of data to improve a learner, several works [47], [26] considered combining semi-supervised learning and active learning in object detection. Rhee et al. [47] proposed ASSL for pedestrian detection, where SSL method provides the incremental improvement of semi-supervised detection performance by combining the concept of diversity imported from AL methods. In [26], Mi et al. extended the teacher-student framework to an iterative version, called ActiveTeacher, where the label set is partially

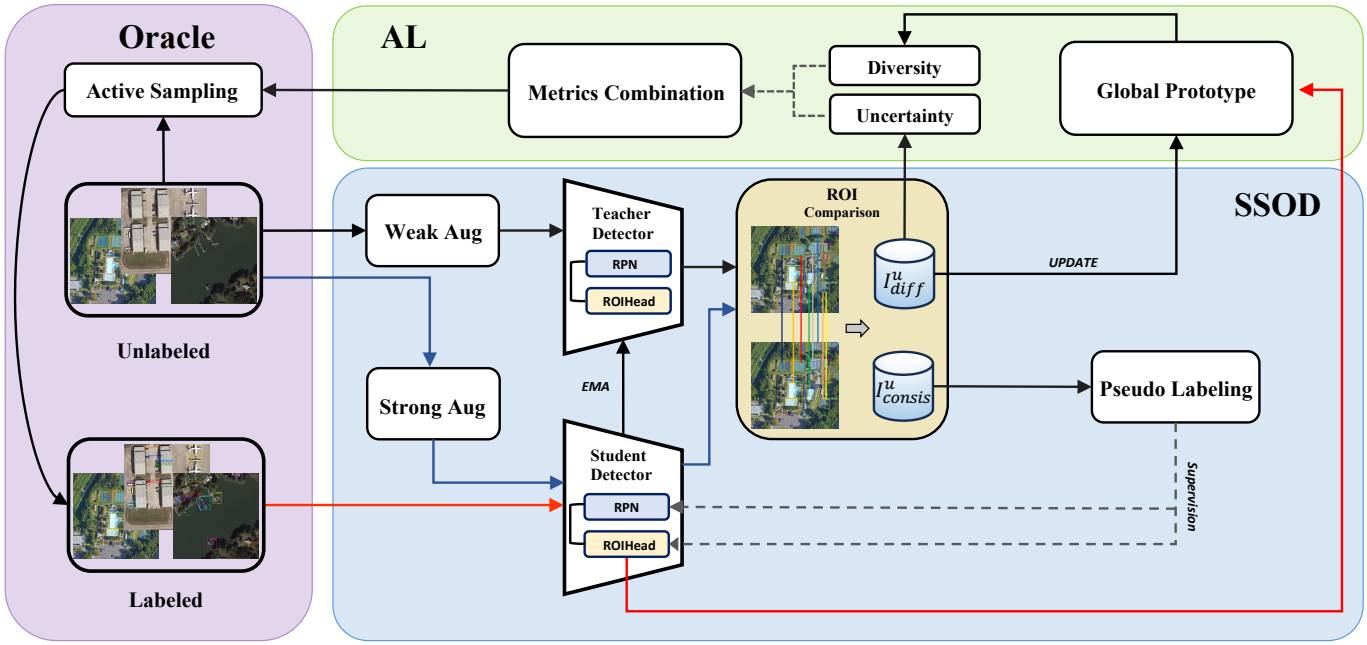


Fig. 2. Overview of the proposed SSOD-AL framework. There are mainly three stages in our method: ***Semi-Supervised Object Detection(SSOD)***: Using limited label set to initialize the parameters of Teacher-Student framework, where Teacher is responsible for generating pseudo-labels to Student while Student is trained with both ground-truth and pseudo-labels. ***Active Learning(AL)***: Select the top-N valuable samples for labeling according to uncertainty and diversity metrics, which are based on RoI Comparison and Category Prototype, respectively. ***Label set Augmentation(Oracle)***: Using the active selected samples to augment the label set, then train the Teacher-Student framework iteratively by repeating the preceding procedures.

initialized and gradually augmented by evaluating three key factors of unlabeled examples, including difficulty, information and diversity. ActiveTeacher [26] is the first attempt of studying data initialization in teacher-student based semi-supervised object detection (SSOD)

III. Proposed Methodology

In this section, we elaborate on the designs of our proposed semi-supervised active learning method in remote sensing object detection, as shown in Fig. 2. Expanding upon the conventional teacher-student framework, we have introduced an iterative structure. Initially, the labeled remote-sensing images constitute a small fraction of the dataset. However, by leveraging active learning, the labeled image set undergoes continuous expansion throughout the iterative process.

A. Semi-Supervised Object Detection

Before introducing the method of semi-supervised object detection(SSOD), we first provide a detailed explanation of the definition of semi-supervised learning. Given a set of labeled data $D_l = \left\{\left(x_i^l, y_i^l\right)\right\}_{i=1}^{N_l}$ and a set of unlabeled data $D_u = \left\{\left(x_j^u\right)\right\}_{j=1}^{N_u}$, where y_i^l denotes the label of image x_i^l, N_l and N_u are the total number of labeled and unlabeled data, respectively. The purpose of semi-supervised learning is to train a high-performance model based on both D_l and D_u.

Similar to previous semi-supervised learning methods [15], [44], [26], our SSOD method also includes two detection networks, called Teacher and Student, as shown in Fig 2. The Teacher-Student network takes the unlabeled image x_j^u as input and pass it through the weak augmentation module T_w and strong augmentation module T_s to obtain input for the Teacher detector M_t and the Student detector M_s, respectively. The Teacher detector is responsible for producing pseudo-labels by the weak augmentation image $T_w\left(x_j^u\right)$, while the Student detector outputs prediction results $M_s\left(T_s\left(x_j^u\right)\right)$ generated by the strong augmentation image $T_s\left(x_j^u\right)$. Detection loss of unlabeled images $\mathcal{L}_{det}^{unsup}$ is defined to minimize the differences between prediction results $M_s\left(T_s\left(x_j^u\right)\right)$ and pseudo labels. Simultaneously, we also feed labeled images x_i^l into the Student detector M_s to obtain prediction results and define a loss function $\mathcal{L}_{det}^{sup}$ to minimize the distance between $M_s\left(x_i^l\right)$ and the ground truth. The total loss of Teacher-Student network is formulated in Equation (1).

$$\mathcal{L}_{det}^{ts}\left(\mathcal{D}_l, \mathcal{D}_u\right) = \mathcal{L}_{det}^{sup}\left(\mathcal{D}_l\right) + \lambda_u \mathcal{L}_{det}^{unsup}\left(\mathcal{D}_u\right) \tag{1}$$

where λ_u is a hyper-parameter to balance the detection losses of labeled and unlabeled images.

Regarding the task of object detection, $\mathcal{L}_{det}^{sup}$ include classification loss $\mathcal{L}_{det}^{cls}$ of RPN and ROI head, and the one for bounding box regression $\mathcal{L}_{det}^{loc}$. Then, $\mathcal{L}_{det}^{sup}$ is defined as

$$\mathcal{L}_{det}^{sup} = \frac{1}{N_l}\sum_{i=1}^{N_l}\left(\mathcal{L}_{det}^{cls}\left(x_i^l, y_i^{cls}\right) + \mathcal{L}_{det}^{loc}\left(x_i^l, y_i^{loc}\right)\right) \tag{2}$$

where $\mathcal{L}_{det}^{cls}$ and $\mathcal{L}_{det}^{loc}$ are defined as

$$\mathcal{L}_{det}^{cls}\left(x_i^l, y_i^{cls}\right) = \mathcal{L}_{rpn}^{cls}\left(x_i^l, y_i^{cls}\right) + \mathcal{L}_{roi}^{cls}\left(x_i^l, y_i^{loc}\right) \tag{3}$$

$$\mathcal{L}_{det}^{loc}\left(x_i^l, y_i^{loc}\right) = \sum_{c\in\{x,y,h,w\}} \text{Smooth}_{L1}\left(t_c^i, y_c^i\right) \tag{4}$$

Algorithm 1 Semi-supervised Active Learning based on RoI Comparison and Category Prototype(SSOD-AL)

Input: Labeled Dataset $D_l^0 = \{(x_l^0, y_l^0)\}$, Unlabeled Dataset $D_u^0 = \{(x_u^0)\}$, Student Detector M_s^0, Teacher Detector M_t^0, RoI Comparison Module R, Object Category Set C, Maximum Iteration K.
Outpt: The ultimate optimal Teacher Detector M_t^K
for all $(x_l^0, y_l^0) \in D_l^0$ and $x_u^0 \in D_u^0$ **do**
  Initialize the parameters of M_s^0 and M_t^0
end for
for all $n = 1, ...K$ **do**
  for each $(x_l^{n-1}, y_l^{n-1}) \in D_l^{n-1}$ **do**
    Extract $f_{i,j}^{gt}$ of ground-truth proposals by using M_s^{n-1}
    for $k \in C$ **do**
      Compute local prototype v_k according to Eq.(12)
      Update global prototype g_k according to Eq.(14)
    end for
  end for
  for each $x_u^{n-1} \in D_u^{n-1}$ **do**
    Using R to compare RoIs generated by M_s^{n-1} and M_t^{n-1}, dividing x_u^{n-1} into I_{diff}^u and I_{consis}^u;
    if $x_u^{n-1} \in I_{consis}^u$ **then**
      Compute KL divergence by using R according to Eq.(8)
    else if $x_u^{n-1} \in I_{diff}^u$ **then**
      Calculate the uncertainty metric S_{unc}^{n-1} by Eq.(10)
    end if
  end for
  Sort $x_u^{n-1} \in I_{diff}^u$ by S_{unc}^{n-1} as I_{sorted}^u
  for each $x_u^{n-1} \in I_{sorted}^u$ **do**
    Calculate the diversity metric S_{div}^{n-1} by Eq. (16) and integrate it with S_{unc}^{n-1}
  end for
  Rank unlabeled samples based on the metrics above.
  Select the top-N samples as x_s^n, annotating them with label y_s^n
  Update labeled set $D_l^n = D_l^{n-1} \cup \{(x_s, y_s)\}$
  Update unlabeled set $D_u^n = D_u^{n-1} - \{(x_s)\}$
  for all $(x_l^n, y_l^n) \in D_l^n$ and $x_u^n \in D_u^n$ **do**
    Update M_s^n according to Eq.(9)
    Update M_t^n according to Eq.(7)
  end for
end for

where y_i^{cls} and y_i^{loc} are classification labels and bounding box labels, t_c^i refers to the c-th coordinate of the output image xi. In terms of $\mathcal{L}_{det}^{loc}$, we use the smooth L-1 loss for the bounding box regression:

$$\mathrm{Smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{5}$$

As for $\mathcal{L}_{det}^{sup}$, we use only predicted pseudo-labels, formulated as

$$\mathcal{L}_{det}^{unsup} = \frac{1}{N_u} \sum_{i=1}^{N_u} \mathcal{L}_{det}^{cls}\left(x_i^u, \hat{y}_i^{cls}\right) \tag{6}$$

where $\mathcal{L}_{det}^{cls}$ is the same as Eq. (3), and $\hat{y}_i^{cls}$ is the pseudo-labels generated by the Teacher network.

To address issues related to class imbalance and over-fitting, we adopted a strategy similar to previous research [15], [26] by freezing the optimization of Teacher network during semi-supervised training. Instead, we updated the its parameters using *Exponential Moving Average* (EMA) [13] based on those of the Student network:

$$\theta_t^i = \alpha\theta_t^i + (1-\alpha)\theta_s^i \tag{7}$$

where θ_t and θ_s represent the parameters of the Teacher and Student networks, respectively, and i denotes the i-th round of training, while α is the hyper-parameter that determines the rate of parameter update, typically close to 1. Additionally, we employ non-maximum suppression (NMS) [48] to filter out duplicate pseudo-labels, and set a confidence threshold to filter out low-confidence pseudo-labels, thereby ensuring the quality of generated pseudo-labels.

B. Active Learning Sampling

The aim of active learning is to choose the most valuable images for annotation, enabling the training of a high-performing network with only a limited number of labeled images. This is especially crucial in remote sensing images, where annotation costs are exceedingly high. In our method, we extend the conventional Teacher-Student network into an iterative one, employing the concept of active learning to semi-supervised object detection. To begin with, we initialize a small subset of labeled images as the label set and introduce a RoI Comparison Module, which compares the RoIs generated by the teacher and student detectors. We also propose a novel active sampling strategy for selecting the most valuable images for annotation and update the label set after each round of semi-supervised training. Through this modification, limited annotations can be optimally utilized by active sampling, resulting in improved quality of pseudo-labels. Similar to previous active learning works, we also select samples for annotations based on image uncertainty and class diversity, introducing an uncertainty selection model assisted by the ***RoI Comparison Module(RoICM)*** and a diversity selection model based on the *Global Class Prototype*, respectively.

1) *Uncertainty selection model assisted by the RoI Comparison Module(RoICM)*: We measure the uncertainty of selected images by analyzing the differences between the proposal bounding boxes predicted by the Teacher and Student network. Given an unlabeled image x_i^u, we first input it into both the Teacher and Student networks to obtain their respective predictions of RoIs. These RoIs are then fed into the ***RoICM*** for comparison. ***RoICM*** employs a set of comparison rules that ensure the highest degree of accuracy and precision in the comparison process:

a. If the RoIs predicted by both the Teacher and Student networks have *consistent* classes, we add x_i^u to the set of images with consistent class predictions I_{consis}^u. We deem the RoIs of $x_i^u \in I_{consis}^u$ predicted by the Teacher network to be reliable and proceed to pseudo-label the

image, allowing to feed it into the Student network for the next stage of training.

b. If the RoIs predicted by both the Teacher and Student networks have *different* classes, we add x_i^u to the set of images with different class predictions I_{diff}^u. This decision is predicated on the fact that $x_i^u \in I_{diff}^u$ is considered to have a higher level of uncertainty for the detection network, thereby increasing its overall annotation value. Thus, they should be included in the active learning selection sequence for human labeling, ensuring that the network remains highly accurate and reliable.

Assisted by *RoICM*, for $x_i^u \in I_{consis}^u$, we calculate the *KL divergence* based on the predicted class distributions by the Teacher and Student networks:

$$D_{kl} = \frac{1}{n_b^i} \sum_{j=1}^{n_b^i} \sum_{k=1}^{N_c} p_t^i \left(b_j, c_k\right) \frac{\log p_t^i \left(b_j, c_k\right)}{\log p_s^i \left(b_j, c_k\right)} \tag{8}$$

where n_b^i represents the number of proposal bounding boxes generated by the Teacher network after NMS and confidence threshold filtering, N_c is the number of instance categories and $p^i \left(b_j, c_k\right)$ is the prediction probability of the k-th category by the network. $b_j^{'}$ and b_j are the corresponding bounding boxes predicted by student and teacher network respectively. The metric of KL divergence, which measures the difference between the probability distributions of the Teacher and Student networks, is a key factor in determining the reliability of the generated pseudo-labels. A lower value of KL divergence indicates a higher degree of consistency between the two networks, resulting in more dependable and trustworthy pseudo-labels produced by the Teacher network. Thus we assign a higher weight to these predictions when calculating the total loss as follows:

$$\mathcal{L}_{det}^{ts} \left(\mathcal{D}_l, \mathcal{D}_u\right) = \mathcal{L}_{det}^{sup} \left(\mathcal{D}_l\right) + \exp(-D_{kl}) \cdot \lambda_u \cdot \mathcal{L}_{det}^{unsup} \left(\mathcal{D}_u\right) \tag{9}$$

Instead, for $x_i^u \in I_{diff}^u$ that have a higher level of uncertainty and annotation value, we measure the degree of uncertainty based on the predicted category distribution of the Teacher network for each instance of x_i^u. The metric for uncertainty is calculated as follows:

$$\cdot S_{unc}^i = -\frac{1}{n_b^i} \sum_{j=1}^{n_b^i} \sum_{k=1}^{N_c} p_t^i \left(b_j, c_k\right) \log p_t^i \left(b_j, c_k\right) \tag{10}$$

Based on (10), we can determine whether an image has higher annotation value by assessing its uncertainty level based on the predictions of the Teacher network.

2) *Diversity selection model based on the Global Class Prototype*: We have established a global prototype for each category, which serves as the basis for ensuring the diversity of the selected image categories. More specifically, given a labeled image x_i^l, we obtain a set of RoI features F_i^{gt} from the ground-truth bounding boxes generated by the RoI Head of the Student Detector in each training stage:

$$\mathrm{F}_i^{gt} = \left\{ \left(f_{i,j}^{gt}, y_{i,j}^{gt}\right) \right\} \tag{11}$$

where $f_{i,j}^{gt}$ denotes the RoI feature of the j-th ground-truth bounding box, $y_{i,j}^{gt} \in C$ is its class label while C is the set of total instance categories. We obtain a local prototype for each class by calculating the average of the RoI features by class:

$$v_k = \begin{cases} \frac{\sum_{i,j} f_{i,j}^{gt} \mathbb{1}\left(y_{i,j}^{gt}=k\right)}{\sum_{i,j} \mathbb{1}\left(y_{i,j}^{gt}=k\right)} & \sum_i \mathbb{1}\left(y_{i,j}^{gt}=k\right) > 0 \\ \mathbf{0} & \sum_i \mathbb{1}\left(y_{i,j}^{gt}=k\right) = 0 \end{cases} \tag{12}$$

where v_k is the local prototype of k-th class in C, $\mathbf{0}$ denotes the zero vector and $\mathbb{1}\left(y_{i,j}^{gt}=k\right)$ is defined as follows:

$$\mathbb{1}\left(y_{i,j}^{gt}=k\right) = \begin{cases} 1 & \text{if } y_{i,j}^{gt}=k \\ 0 & \text{if } y_{i,j}^{gt} \neq k \end{cases} \tag{13}$$

The global class prototype is updated using the EMA algorithm [13], with the local class prototype serving as a reference:

$$g_k = \alpha g_k + (1-\alpha) v_k \tag{14}$$

where g_k is the global prototype of k-th class in C, α denotes the hyper-parameter that is typically closed to 1. By applying semi-supervised training on a small initial set of labeled data, we are able to obtain the initial global class prototype, which serves as a foundation for ensuring the diversity and effectiveness of the subsequent active learning process.

For unlabeled images, we only adopt $x^u \in I_{diff}^u$ to measure their diversity and update the global category prototype, which is due to the fact that this part of the images already has a certain level of uncertainty and thus is more representative in unlabeled set. First, to make the updating process of the global category prototype smoother, we sort $x_i^u \in I_{diff}^u$ from smallest to largest uncertainty values. As described in Section 2.2.1, the weak augmented unlabeled image $T_w\left(x_j^u\right)$ is first fed into the Teacher network, which subsequently generates a set of proposal bounding boxes b_j through NMS [48] and confidence threshold filtering. b_j is then input into the RoI head of the Student network to obtain its RoI feature $f_{i,j}^{pgt}$. To measure the similarity between $f_{i,j}^{pgt}$ and g_k, we adopt cosine similarity as a metric:

$$\mathrm{sim}\left(f_{i,j}^{pgt}, g_k\right) = \frac{f_{i,j}^{pgt} \left(g_k\right)^T}{\|g_k\| \cdot \left\|f_{i,j}^{pgt}\right\|} \tag{15}$$

For each RoI feature $f_{i,j}^{pgt}$, the class c in the global class prototype with the highest similarity score can be identified, denoted as $\max_{k \in N_c} \mathrm{sim}\left(f_{i,j}^{pgt}, g_k\right)$. If the similarity score falls below a certain threshold s, it suggests that the target object might belong to a novel class that is not yet present in the global class prototype. Consequently, the global class prototype must be updated using equations (12) and (14).

On the contrary, to ensure the diversity of selected samples, we need to suppress instances with high similarity to the global class prototype. Therefore, we can calculate the diversity score S_{div}^i of the unlabeled image $x_i^u \in I_{diff}^u$. S_{div}^i represents the average similarity between x_i^u and all other unlabeled images in the feature space. The higher the diversity score, the more dissimilar the image is to other images, indicating that it may

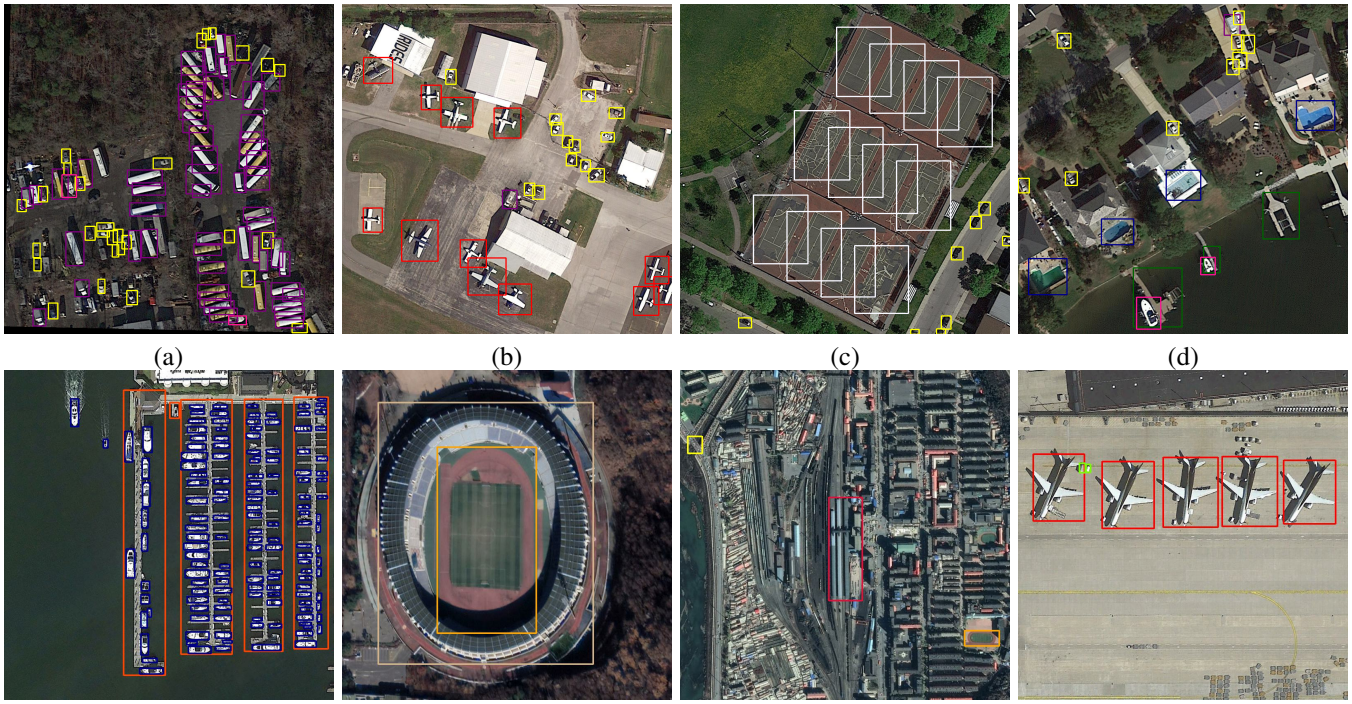


Fig. 3. Visualization of representative labeled samples in DOTA and DIOR datasets. Images (a)-(d) are labeled samples from DOTA dataset, while (e)-(h) are labeled samples from DIOR dataset.

represent a novel class or a new aspect of a known class. The diversity score S^i_{div} can be calculated as follows:

$$S^i_{div} = 1 - \frac{1}{n^i_b} \sum_{j=1}^{n^i_b} \max_{k \in N_c} \text{sim}\left(f^{pgt}_{i,j}, g_k\right) \tag{16}$$

Finally, we combine the two metrics as our ultimate sampling metrics. Before combining, we first normalize the metrics separately due to the possible deviations in the range of values:

$$\widehat{S^i_s} = \frac{S^i_s}{S^{\max}_s} \tag{17}$$

where $s \in \{unc, div\}$ denote our two sampling strategies. After that, we use *L-p* normalization to obtain the final selection score of the unlabeled image x^u_i:

$$S^i_{sel} = \sqrt[p]{(S^i_{unc})^p + \left(S^i_{div}\right)^p} \tag{18}$$

In general, we use L-1 to combine these two metrics. With the reference of S^i_{sel}, the most informative samples are selected so that the limited labeled samples fed into the Teacher-Student framework in each iteration are augmented with high quality, further improving the performance of the detection head.

IV. Experiment And Analysis

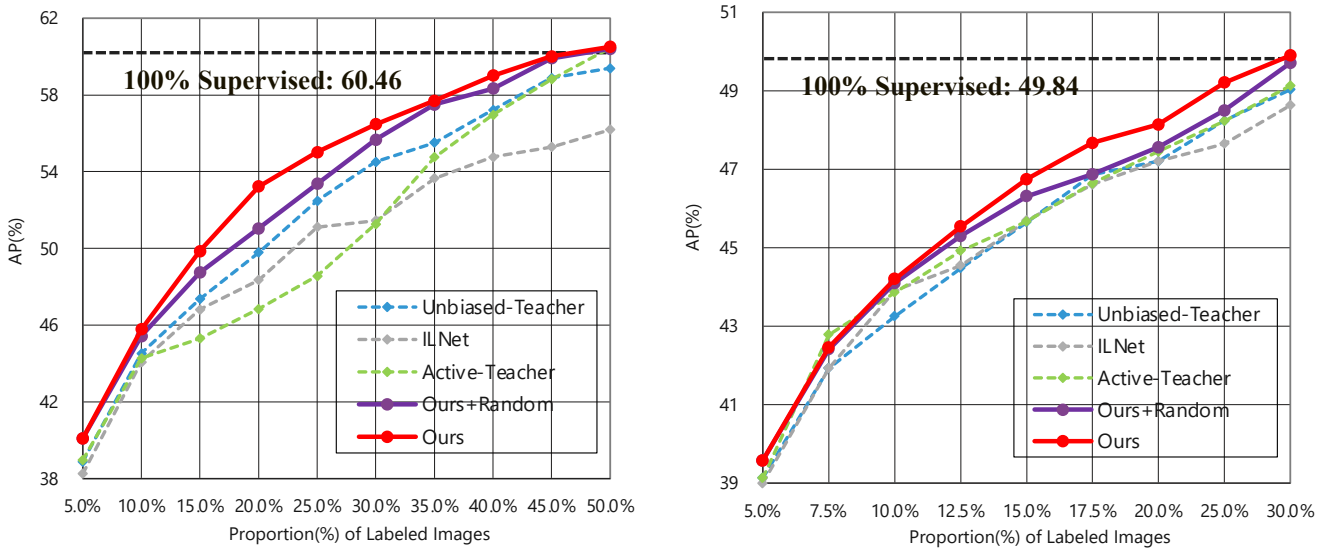
In this section, comprehensive experiment were conducted to demonstrate the effectiveness and superiority of the proposed SSOD-AT. We first introduce the datasets and evaluation metrics, and then, quantitative and qualitative results of our method on three public datasets are shown and analyzed, and compared to other state-of-the-art methods. Finally, ablation studies on our RoI comparison module and active learning strategies are covered and reviewed.

A. Datasets

To extensively evaluate the proposed framework, two representative and public datasets in remote-sensing images, known as DOTA [49] and DIOR [50] were employed in our experiments.

1) DOTA Dataset: DOTA [49] is a large annotated object dataset with a wide variety of categories in Earth Vision, which contains 2806 aerial images from different sensors and platforms with crowd sourcing and 188,282 instances, covered by 15 common object categories, including *plane (PL), baseball diamond (BD), bridge (BR), ground field track (GFT), 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)*. As same as the formal guide, 2/3 and 1/3 of the original images are randomly selected for training and testing, respectively. Furthermore, we crop a series of 1024 × 1024 patches from the original images with a stride of 824. In our experiments, we focus on the task of detection with horizontal bounding boxes (**HBB** for short), evaluating our method with HBB ground truths.

2) DIOR Dataset: DIOR [50] is a large scale, publicly available benchmark for object detection in optical remote-sensing images, which consists of 23,463 images and 192,472 instances that are manually labeled with axis-aligned bounding boxes, covering 20 object categories. These 20 object classes are *airplane, airport, baseball field, basketball court, bridge,*



(a) (b)

Fig. 4. Detection results of the different algorithms on the two remote-sensing datasets. (a) DOTA. (b) DIOR

TABLE I
COMPARISON BETWEEN DOTA AND DIOR.

Dataset	Category	Image quantity	BBox quantity	Avg. BBox quantity
DOTA	15	2,806	188,282	67.10
DIOR	20	23,463	192,472	8.20

chimney, dam, expressway service area, expressway toll station, harbor, golf course, ground track field, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and wind mill. We divide all images of DIOR into three subsets, training set and test set, with a ratio of 2:1, respectively. DIOR has some remarkable characteristics that can be used to evaluate the merits of our approach in remote-sensing images, including large scale, a large range of object size variations, rich image variations, high inter-class similarity and intra-class diversity.

The comparison between DOTA and DIOR is demonstrated in Table I. *BBox* is short for bounding boxes, while *Avg. BBox quantity* indicates average bounding box quantity per image. Note that for the average number of instances per image, DOTA surpasses DIOR hugely, suggesting that most of the samples in DOTA may be typically more complicated compared to DIOR.

B. Design of the Experiments

During our experiments, similar to the previous works in SSOD [15], [26], the training set of datasets are further divided into the labeled set and the unlabeled set. The labeled set first randomly selects 5% labeled images from training set of DOTA and 2.5% of DIOR, while remaining images in the training set are used to construct the unlabeled set. As the iterative training SSOD proceeds, the labeled set gradually augments from actively selected samples to 50%, 30% for the training set of DOTA and DIOR, respectively. For evaluation, we follow the previous works [16], [42], [15], [44], [26] adopt mAP (50:95) [51] as the basic comparison metric in our experiments. We also use paired t-test at 95% significance level with p-value for further comparison. T-test is widely used to verify the difference between two methods, which can suggest which method is better in the whole active learning process [52]. If the p-value is less than or equal to 0.05, it indicates that the two methods have a significant difference, and the larger average accuracy of the method in this comparison is superior to the other method.

As for building our SSOD-AT framework, we adopt Faster-RCNN with ResNet-50 as our basic detection network, as same as the previous successful SSOD efforts [15], [26]. Corresponding to Detectron2 [43], the method of execution and hyper-parameter settings are employed. On the stage of SSOD, we pre-trained the Teacher detector with the limited labeled samples for 4k iterations in all experiments, and the Student detector is initialized by the parameters of the teacher one. Furthermore, we set the whole semi-supervised learning training steps to 50k on 2 GPUs, and we use a threshold $\tau = 0.7$ to filter the pseudo-labels of low quality. The optimizer employed is SGD [53], and at 49,990 and 49,995 iterations, respectively, the learning rate is split by 10 and grows linearly from 0.001 to 0.01 over the first 1k iterations. For the hyperparameters that occur in the equations of our method, we set $\alpha = 0.9996$ for EMA in Eq.(7) and Eq.(14), $\lambda = 4$ for the unsupervised loss in Eq.(6) on all experiments. In terms of s in computing diversity metric, we performed a series of ablation experiments to find the optimal threshold. The batch size for training is set to 32, which consists 16 labeled and 16 unlabeled images via random sampling. For the batch size h of active sampling, we selected the top 5% samples in DOTA and the top 2.5% samples in DIOR based on their respective metrics for human labeling in each iteration. Meanwhile, we conducted parametric analysis

TABLE II
COMPARED WITH STATE-OF-THE-ART METHODS ON THE DOTA DATASET. THE METRIC IS MAP(50:95). "SUPERVISED" REFERS TO THE MODEL FED BY LABELED DATA ONLY. * IS THE ORIGIN SSOD MODEL FED BY OUR ACTIVE SAMPLED DATA. Δ: AP GAIN TO THE SUPERVISED PERFORMANCE

Method	DOTA													
	L+5%	Δ	L+10%	Δ	L+15%	Δ	L+20%	Δ	L+25%	Δ	L+35%	Δ	L+45%	Δ
Supervised	26.62	+0.00	32.48	+0.00	36.71	+0.00	40.84	+0.00	43.83	+0.00	47.41	+0.00	52.16	+0.00
CALD	27.13	+0.51	35.46	+2.98	40.01	+3.30	43.09	+2.25	45.47	+1.64	50.83	+3.42	54.42	+2.26
Unbiased-Teacher	44.53	+17.91	47.36	+14.88	49.79	+13.08	52.49	+11.65	54.52	+10.69	57.22	+9.81	59.40	+7.24
ILNet	44.08	+17.46	46.81	+14.33	48.35	+11.64	51.12	+10.28	51.46	+7.63	54.77	+7.36	56.19	+4.03
Unbiased-Teacher*	44.31	+17.69	49.21	+16.73	52.19	+15.48	53.95	+13.11	55.55	+11.72	57.65	+10.24	59.88	+7.72
ILNet*	43.82	+17.20	46.97	+14.49	49.23	+12.52	50.64	+9.80	52.78	+8.95	54.61	+7.20	56.55	+4.39
Active-Teacher	44.27	+17.65	45.31	+12.83	46.84	+10.13	48.57	+7.73	51.27	+7.44	56.97	+9.56	60.46	+8.30
SSOD-AT(Random)	45.43	+18.81	48.75	+16.27	51.04	+14.33	53.37	+12.53	55.67	+11.84	58.34	+10.93	60.41	+8.25
SSOD-AT(Ours)	45.78	+19.16	49.86	+17.38	53.23	+16.52	55.02	+14.18	56.47	+12.64	59.02	+11.61	60.68	+8.35

TABLE III
PROPOSED SSOD-AT METHOD VERSUS THE COMPARISON METHODS BASED ON PAIRED T-TESTS AT 95% SIGNIFICANT LEVEL ON THE DOTA DATASET

Label Proportion	Ours vs CALD	Ours vs Unbiased-Teacher	Ours vs ILNet	Ours vs Unbiased-Teacher*	Ours vs ILNet*	Ours vs Active-Teacher
	P-value	P-value	P-value	P-value	P-value	P-value
L+5%	0.06	0.01	0.025	0.06	0.020	0.09
L+10%	0.016	0.06	0.036	0.044	0.021	0.16
L+15%	4.03e-03	0.03	0.03	7.89e-03	0.013	0.07
L+20%	1.20e-03	6.79e-03	7.14e-03	1.23e-03	4.31e-03	0.025
L+25%	3.94e-04	1.66e-03	2.24e-03	2.26e-04	8.27e-04	7.15e-03
L+35%	9.12e-05	8.28e-05	9.38e-05	1.06e-05	4.83e-05	1.62e-03
L+45%	3.78e-05	2.54e-05	3.62e-06	1.26e-06	1.63e-06	2.09e-03

experiments to establish the optimality of our chosen batch size. We recorded the experimental results for each round until the model performed at the level of the fully supervised model. For all comparative algorithms [25], [15], [44], [26], we seamlessly adapted them from their original application on natural scenes to remote sensing scenes while preserving their inherent configurations. In the case of SSOD-based methods [15], [44], [26], we ensured fairness by setting their semi-supervised learning training steps to 50k, consistent with our own approach. This decision was based on the observation that they commonly achieve convergence at this juncture.

C. Comparative Experiments With State-of-the-Art Methods

In this section, we compare the proposed SSOD-AT with other state-of-the-art methods on two popular aerial object detection datasets: DIOR and DOTA. The methods we select for comparison [25], [15], [44], [26] are all the latest proposed AL, SSOD and SSAL methods. Among them, CALD [25] is a detection-specific active learning method, which considers the consistency of both bounding box and predicted class distribution when augmentation is applied to overcome the challenges brought by inconsistencies between classification and detection. Unbiased-Teacher [15] first apply EMA [13] and focal loss [43] to address the pseudo-label over-fitting problem in teacher-student learning. In addition, both ILnet [44] and Active-Teacher [26] are based on Unbiased-Teacher [15], where ILnet [44] introduces an additional classification task for bounding box localization to improve the filtering of the predicted bounding boxes, while Active-Teacher [26] first introduces a novel active sampling strategy to augment the label set participating in each training iteration.

Furthermore, we extended the ILnet and Unbiased-Teacher methods, denoting the augmented versions as ILnet* and Unbiased-Teacher* respectively. Building upon the original algorithms, we replaced the involvement of annotated samples during training from random sampling with the proposed active sampling strategy in SSOD-AT. This deliberate modification was undertaken to empirically assess the efficacy of the active sampling approach introduced in SSOD-AT.

1) Results on DOTA: The proposed SSOD-AT is evaluated on DOTA and compared to other representative AL, SSOD and SSAL methods in Table II, Table III and Fig. 4(a), all of which are based on Faster-RCNN ResNet-50 horizontal detectors. As demonstrated by the results of **SSOD-AT(Random)** in Table II, even trained with randomly selected samples for human labeling, our method consistently outperforms the

TABLE IV
COMPARED WITH STATE-OF-THE-ART METHODS ON THE DIOR DATASET. THE METRIC IS MAP(50:95). "SUPERVISED" REFERS TO THE MODEL FED BY LABELED DATA ONLY. * IS THE ORIGIN SSOD MODEL FED BY OUR ACTIVE SAMPLED DATA. Δ: AP GAIN TO THE SUPERVISED PERFORMANCE

Method	DIOR													
	L+5%	Δ	L+10%	Δ	L+15%	Δ	L+20%	Δ	L+25%	Δ	L+35%	Δ	L+45%	Δ
Supervised	22.87	+0.00	25.28	+0.00	27.03	+0.00	29.34	+0.00	30.45	+0.00	32.49	+0.00	34.61	+0.00
CALD	24.10	+1.23	26.96	+1.68	28.85	+1.82	30.93	+1.59	32.21	+1.76	34.57	+2.08	36.45	+1.84
Unbiased-Teacher	43.25	+20.38	44.48	+19.20	45.65	+18.62	46.86	+17.52	47.20	+16.75	48.24	+15.75	49.03	+14.42
ILNet	43.91	+21.04	44.55	+19.27	45.69	+18.66	46.61	+17.27	47.20	+16.75	47.65	+15.16	48.63	+14.02
Unbiased-Teacher*	43.99	+21.12	45.22	+19.94	46.05	+19.02	47.07	+17.73	47.53	+17.08	48.36	+15.87	49.05	+14.44
ILNet*	43.77	+20.90	45.07	+19.79	46.06	+19.03	46.90	+17.56	47.38	+16.93	47.88	+15.39	48.61	+14.00
Active-Teacher	43.86	+20.99	44.92	+19.64	45.67	+18.64	46.64	+17.30	47.48	+17.00	48.23	+15.74	49.13	+14.52
SSOD-AT(Random)	44.10	+21.23	45.30	+20.02	46.31	+19.28	46.87	+17.53	47.56	+17.11	48.50	+16.01	49.71	+15.10
SSOD-AT(Ours)	44.20	+21.33	45.54	+20.26	46.75	+19.72	47.67	+18.33	48.14	+17.69	49.21	+16.72	49.90	+15.29

TABLE V
PROPOSED SSOD-AT METHOD VERSUS THE COMPARISON METHODS BASED ON PAIRED T-TESTS AT 95% SIGNIFICANT LEVEL ON THE DOTA DATASET

Label Proportion	Ours vs CALD	Ours vs Unbiased-Teacher	Ours vs ILNet	Ours vs Unbiased-Teacher*	Ours vs ILNet*	Ours vs Active-Teacher
	P-value	P-value	P-value	P-value	P-value	P-value
L+5%	4.48e-03	0.053	0.071	0.041	7.68e-03	0.58
L+10%	7.61e-04	0.016	0.026	7.58e-03	6.95e-04	0.28
L+15%	1.25e-04	3.92e-03	9.07e-03	6.78e-03	2.77e-04	0.13
L+20%	2.54e-05	7.48e-04	2.49e-03	1.65e-03	1.07e-04	0.052
L+25%	5.71e-06	1.28e-04	5.41e-04	3.70e-04	2.22e-05	0.021
L+35%	1.70e-06	2.09e-05	4.27e-04	2.07e-04	2.26e-04	7.51e-03
L+45%	6.35e-07	3.17e-06	1.12e-04	7.37e-05	1.24e-04	2.61e-03

comparison methods at each labeling ratio. This highlights the effectiveness of our strategy, which ensures the quality of pseudo-labels generated by the Teacher-Detector through the introduced RoICM, making our method well-suited for the challenges of remote sensing images. Furthermore, by employing our active learning sampling strategy to further enhance the informativeness of the labeled samples, the performance of our proposed method exhibits significant improvement, as evidenced by the results of **SSOD-AT** in Table II and Table III. The performance of SSOD-AT has advantage than the other comparison methods at many iterations in the whole active-learning process on DOTA dataset. Similarly, Unbiased-Teacher* and ILNet*, obtained from training samples sampled using our active sampling strategy, also show improved performance over their original algorithms, further demonstrating the efficiency of our active learning strategy in selecting high-quality samples for human labeling. Moreover, as illustrated in Fig. 4(a), our method achieves comparable detection capability to the fully supervised model with only 50.0% of the samples labeled, thereby substantially reducing the number of labeled samples required for training.

2) Results on DIOR: Fig. 4(b), Table IV and Table V present a comparative analysis of the proposed method against other state-of-the-art algorithms on the DIOR dataset. Unlike the results observed on DOTA, the performance variations among the algorithms on DIOR exhibit a slightly smoother pattern, as depicted in Fig. 4(b). This characteristic can be attributed to the sparser distribution of objects in individual images of DIOR compared to DOTA. In Table IV and V, the CALD algorithm, lacking semi-supervised learning, displays inferior performance on DIOR compared to other SSOD and SSAL methods. Conversely, SSOD-AT continues to exhibit superior performance over other state-of-the-art methodologies. From a framework design perspective, as evident in **SSOD-AT(Random)**, SSOD-AT with RoICM still outperforms the two SSOD algorithms, Unbiased-Teacher, and ILNet, even under the premise of randomly selecting labeled samples in each iteration. On the other hand, when considering the quality of training samples in each iteration, SSOD-AT is slightly outperformed by the state-of-the-art SSAL method, Active-Teacher, which capitalizes on our powerful active learning sampling strategy. Meanwhile, the performance of Unbiased-teacher* and ILNet* on DIOR continues to benefit from the original algorithms, further affirming the advantages of our active sampling algorithm. Notably, none of the compared algorithms achieves the performance level of the fully supervised

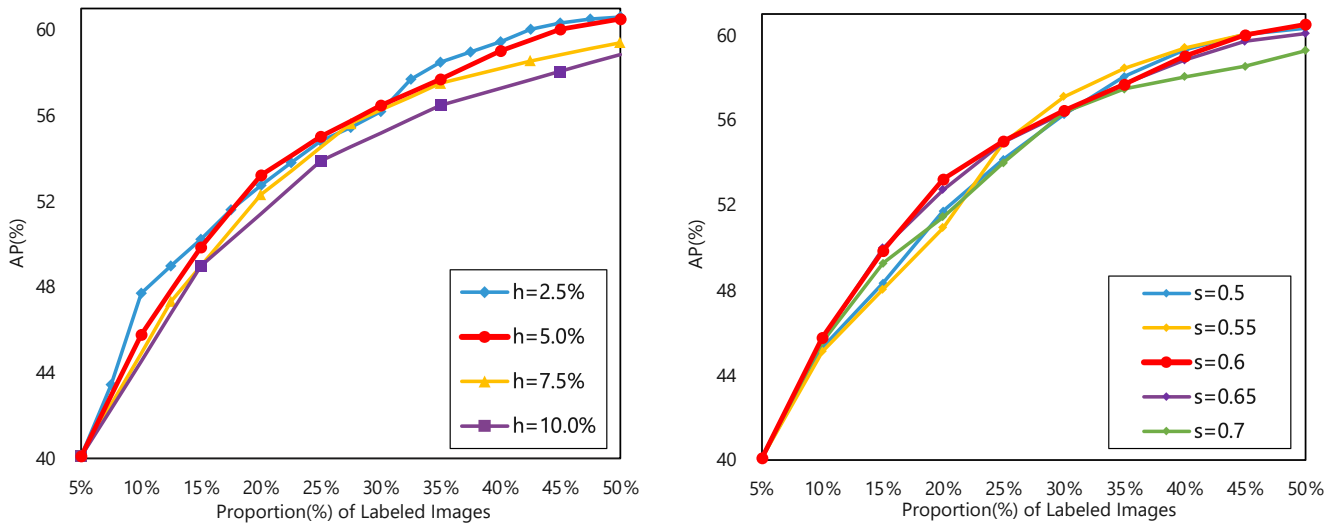


Fig. 5. Results of parameter analysis on DOTA dataset. (a) batch size h analysis. (b) hyperparameter s analysis

detection head. However, SSOD-AT attains this capability by labeling a mere 30% of the samples in the DIOR dataset, as illustrated in Fig. 4(b). In practical applications, this substantial reduction in the cost of labeling for remote sensing images underscores the advantages of our method.

Overall, compared to other state-of-the-art AL, SSOD and SSAL methods tailored for natural images, our method not only optimizes the framework of semi-supervised learning for the characteristics of multi-scale and multi-category targets in remote sensing images, but also proposes a more effective active learning strategy, which enables our method to achieve much higher than expected detection performance on remote sensing images.

D. Ablation Study of Proposed Method

To elucidate the impact of the changes on the overall performance, which systematically removing specific elements from the model architecture or altering its configuration, in this section, we present a detailed description of our conducted ablation experiments in the context of our proposed SSOD-AT on DOTA and DIOR datasets. Specifically, we ablate the two main innovations of SSOD-AT, *RoICM* and *Sampling Strategy*. Through meticulous experimentation and rigorous analysis, we aim to discern the essential components and their respective influences on the model's performance.

1) Ablation of RoI Comparison Module: As previously expounded, the RoI Comparison Module(RoICM) constitutes a vital element of SSOD-AT. It ensures the veracity of pseudo-labels by conducting a comparison between the ROIs derived from the teacher and student networks. Moreover, RoICM proficiently filters out unlabeled samples exhibiting heightened uncertainty, thereby enhancing the method's robustness. Hence, we commenced with a meticulous ablation experiment on this pivotal module, as delineated in rows 1 and 2 of Table VI and VII. Within the scope of this section's experimentation, RoICM chiefly fulfills the primary function

TABLE VI
ABLATION STUDY OF PROPOSED SSOD-AT ON DOTA DATASET

Method	RoICM	Sampling Strategy		DOTA		
		Uncertainty	Diversity	10.0%	15.0%	20.0%
Baseline	×	×	×	44.53	47.36	49.79
SSOD-AT	✓	×	×	45.43	48.75	51.04
	✓	✓	×	45.34	49.17	51.87
	✓	✓	✓	**45.87**	**49.43**	**52.57**

TABLE VII
ABLATION STUDY OF PROPOSED SSOD-AT ON DIOR DATASET

Method	RoICM	Sampling Strategy		DIOR		
		Uncertainty	Diversity	10.0%	15.0%	20.0%
Baseline	×	×	×	43.25	45.65	47.20
SSOD-AT	✓	×	×	44.10	46.31	47.56
	✓	✓	×	44.15	46.60	47.88
	✓	✓	✓	**44.20**	**46.75**	**48.14**

described above. Concretely, Table 1 exhibits a comprehensive comparison, wherein the first row represents a conventional semi-supervised teacher-student framework baseline, while the second row introduces our novel RoICM module in conjunction with the baseline, with both training samples randomly extracted from the labeled dataset. Remarkably, the integration of RoICM endows our approach with an enhanced adaptability to target-dense and intricately-structured remote sensing image datasets, yielding a conspicuous and non-negligible enhancement in the model's performance.

2) Ablation of Sampling Strategy: It is well known that uncertainty and diversity are two of the most significant sampling metrics in active learning, and the same has been employed in SSOD-AT. In order to verify the effectiveness of these two metrics respectively, we conducted an ablation experiment with different proportions of labeled data, as shown in rows 3 and 4 of Table VI and VII. It should be noted here that we did not specifically ablate the diversity metrics, since they are

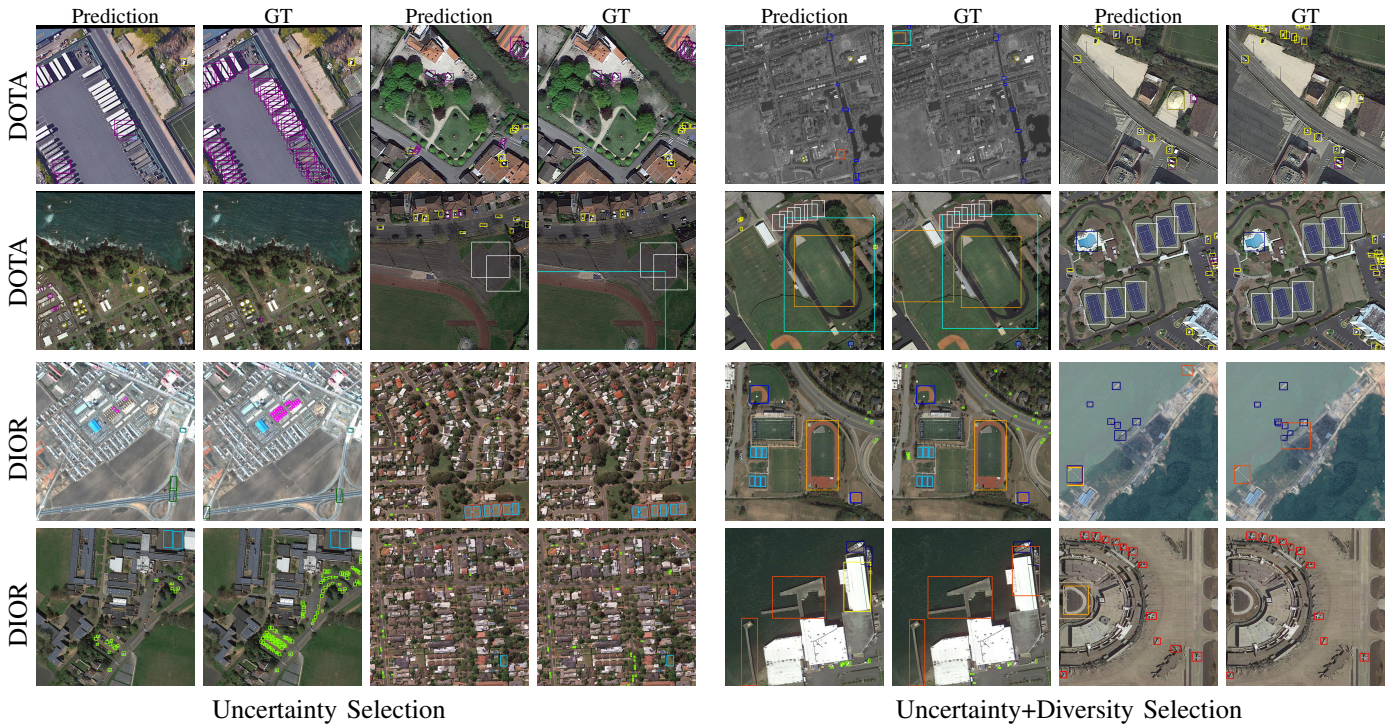


Fig. 6. Visualization of the images with top rank selection priority with different active sampling strategies. The Prediction columns denotes the pseudo-labels predicted by teacher network with 30%(DOTA) and 20%(DIOR) labeled proportions, while the GT columns refers to the corresponding ground-truths.

derived on the basis that the samples are already uncertain to some extent in SSOD-AT, as indicated by Section III-B. Upon scrutinizing rows 3 and 4 of the table, it is evident that the incorporation of both active sampling strategies is beneficial for SSOD. This discernible enhancement stems primarily from the enriched information content of samples under identical label proportions. However, a meticulous comparison between the two active sampling strategies reveals a consistent dominance of the uncertainty metric, derived from the sophisticated RoICM module, in significantly bolstering the model's efficacy across various settings of label proportions. Furthermore, when synergistically amalgamating both metrics for sample selection, the integration of diversity engenders a subtle yet noteworthy amelioration in the model's performance, thus offering compelling validation for the effectiveness of our proposed SSOD-AT methodology.

E. Parameter Analysis

In the pursuit of achieving optimal performance and robustness in SSOD-AT, thorough parameter analysis plays a pivotal role. In this section, we present a comprehensive investigation of parameter analysis in the context of our proposed SSOD-AT on DOTA dataset. Through systematic experimentation and meticulous evaluation, we aim to shed light on the crucial aspects that contribute to the model's performance, including batch size h and hyperparameter α.

1) The analysis of batch size h: Batch size h is a crucial parameter in the active-learning process since it defines the quantity of informative samples are available for human labeling at each iteration. We analyze this parameter by conducting experiments on a candidate set comprising 2.5%, 5.0%, 7.5%, 10.0% of the entire DOTA dataset. Fig. 1 illustrates the efficacy with various batch sizes. Observably, SSOD-AT is sensitive to the batch size h, and a reduced batch size can cause SSOD-AT to converge more rapidly, achieving a higher level of performance with the same number of labeled samples. One possible explanation for this is that a smaller batch size may contain low redundancy information in the query set Q. When the same number of samples is queried, the smaller batch size may contain more information than the larger batch size.

2) Selection of Hyperparameter s: The hyperparameter s is introduced in section III-B and functions as a threshold. If the similarity between the local class prototype v_k of a sample and the global class prototype g_k falls below s, it triggers the need to update the corresponding global class prototype. s completely governs the update process of the g_k which plays a pivotal role in influencing the model's performance. Consequently, conducting comprehensive experiments to ascertain the optimal value of s becomes imperative. Since the similarity usually floats in the interval of 0.5-0.7, we conducted experiments on $s \in \{0.5, 0.55, 0.6, 0.65, 0.7\}$, as shown in Fig. 2. It is evident that the model with $s = \{0.5, 0.55\}$ performs poorly when the labeled samples are scarce, whereas the one with $s = \{0.65, 0.7\}$ exhibits a slower convergence rate. One possible explanation for this phenomenon is as follows: A smaller s implies that g_k is predominantly influenced by a limited number of v_k with lower similarity. As a consequence, g_k may contain an excessive amount of noise information, leading to subpar performance when labeled samples are limited. On the other hand, a larger s indicates that g_k is heavily influenced by numerous v_k with higher similarity. This

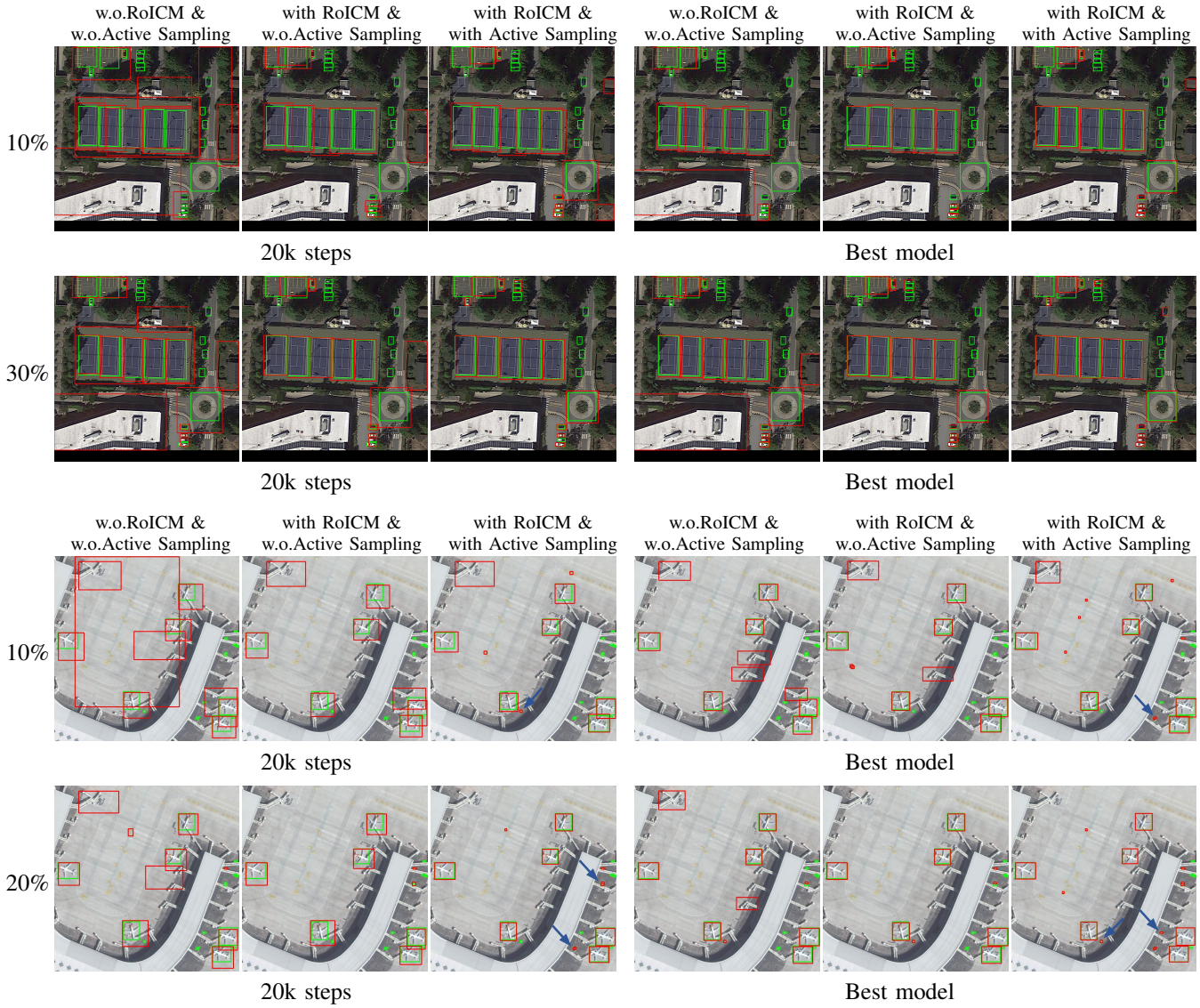


Fig. 7. Visualization of the pseudo-labels predicted by SSOD-AT with and without RoICM and active sampling at different training steps. The green bounding boxes are the ground-truths, while the red ones are pseudo-labels predicted by the teacher network.

hampers the learning of diverse within-class representations in g_k, directly impacting the model's convergence speed. Based on the consideration for achieving the optimal performance and robustness of the model, we have set s = 0.6 for all experiments.

F. Visualization Analysis

1) Visualization of Examples Selected by SSOD-AT Sampling Strategies: In Fig. 6, we visualize the examples selected by active sampling strategies based on 30%(DOTA) and 20%(DIOR) labeled data. It is observable that the uncertainty selection usually selects images with objects that are difficult to detect(e.g., small and occluded objects). Comparing Prediction and GT, we can deserve that a large number of objects are missed or misdetected since the detector is highly uncertain about these images, which also confirms the annotation value of them. However, the samples selected by uncertainty are greatly susceptible to be category imbalanced (e.g., a significant fraction of images containing vehicles). Therefore, it can be seen that the addition of the diversity selection makes more images containing new categories or categories of new form recieve highly attention, which also reflected in the large difference between the category prediction and GT. These images also need expert labeling to improve the generalization ability of SSOD network.

2) Visualization of Effects of Proposed RoICM and Sampling Strategies: We further visualize the pseudo-labels of SSOD-AT with and without RoICM and active sampling with different training steps and label proportions on DOTA and DIOR datasets. As shown in Fig7, although there is still an obvious gap between the qualities of the pseudo-labels and the ground-truth ones, with the assist of RoICM and our active sampling strategies, SSOD-AT still achieves the desired effects. By comparing the RoIs predicted by teacher and student network, RoICM can help the detector successfully filter a great part of unreliable pseudo-labels. Besides, we can

find that our SSOD-AT is also able to detect more small objects in image since the incorporation of our AL query criterion while further improving the prediction accuracy.

V. Conclusion

In this paper, we present a novel semi-supervised active-learning (SSOD-AT) algorithm designed specifically for remote-sensing image object detection. SSOD-AT integrates the RoI Comparison Module(RoICM) and Category Prototype to ensure the reliability of pseudo-labels generated by the Teacher Detector and to effectively select the most valuable images from a vast pool of remote-sensing images for human labeling, thereby constructing a high-quality detection model efficiently. Our proposed method was extensively evaluated on two remote-sensing datasets, where it consistently outperformed state-of-the-art approaches. Remarkably, it achieved performance comparable to that of the fully supervised model using only a fraction of the entire dataset. The detailed experimental analysis provided clear evidence that both uncertainty and diversity play pivotal roles in selecting the most informative samples during the active learning process.

References

[1] M. ElMikaty and T. Stathaki, "Detection of cars in high-resolution aerial images of complex urban environments," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5913–5924, 2017.

[2] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 265–278, 2019.

[3] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[4] C. Li, R. Cong, C. Guo, H. Li, C. Zhang, F. Zheng, and Y. Zhao, "A parallel down-up fusion network for salient object detection in optical remote sensing images," *Neurocomputing*, vol. 415, pp. 411–420, 2020.

[5] W. Ma, N. Li, H. Zhu, L. Jiao, X. Tang, Y. Guo, and B. Hou, "Feature split–merge–enhancement network for remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

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

[7] Z. Dong, M. Wang, Y. Wang, Y. Zhu, and Z. Zhang, "Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2104–2114, 2019.

[8] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu *et al.*, "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 116–130, 2022.

[9] D. Wan, R. Lu, S. Wang, S. Shen, T. Xu, and X. Lang, "Yolo-hr: Improved yolov5 for object detection in high-resolution optical remote sensing images," *Remote Sensing*, vol. 15, no. 3, p. 614, 2023.

[10] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2. Atlanta, 2013, p. 896.

[11] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019.

[12] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[13] 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.

[14] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[15] 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.

[16] 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.

[17] Y. Tang, W. Chen, Y. Luo, and Y. Zhang, "Humble teachers teach better students for semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3132–3141.

[18] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 113–123.

[19] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[20] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in neural information processing systems*, vol. 33, pp. 6256–6268, 2020.

[21] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research*, vol. 2, no. Nov, pp. 45–66, 2001.

[22] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *International conference on machine learning*. PMLR, 2017, pp. 1183–1192.

[23] S. Agarwal, H. Arora, S. Anand, and C. Arora, "Contextual diversity for active learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 137–153.

[24] T. Yuan, F. Wan, M. Fu, J. Liu, S. Xu, X. Ji, and Q. Ye, "Multiple instance active learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5330–5339.

[25] W. Yu, S. Zhu, T. Yang, and C. Chen, "Consistency-based active learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3951–3960.

[26] P. Mi, J. Lin, Y. Zhou, Y. Shen, G. Luo, X. Sun, L. Cao, R. Fu, Q. Xu, and R. Ji, "Active teacher for semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 482–14 491.

[27] 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.

[28] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.

[29] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[30] L. Chen, C. Liu, F. Chang, S. Li, and Z. Nie, "Adaptive multi-level feature fusion and attention-based network for arbitrary-oriented object detection in remote sensing imagery," *Neurocomputing*, vol. 451, pp. 67–80, 2021.

[31] Z. Cui, J. Leng, Y. Liu, T. Zhang, P. Quan, and W. Zhao, "Sknet: Detecting rotated ships as keypoints in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8826–8840, 2021.

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

[33] C. Zhang, B. Xiong, X. Li, and G. Kuang, "Tcd: Task-collaborated detector for oriented objects in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

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

[35] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proceedings of*

the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6728–6736.

[36] M. Gao, Z. Zhang, G. Yu, S. Ö. Arık, L. S. Davis, and T. Pfister, "Consistency-based semi-supervised active learning: Towards minimizing labeling cost," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 510–526.

[37] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5070–5079.

[38] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.

[39] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[40] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[41] A. Kumar, P. Sattigeri, and T. Fletcher, "Semi-supervised learning with gans: Manifold invariance with improved inference," *Advances in neural information processing systems*, vol. 30, 2017.

[42] Q. Zhou, C. Yu, Z. Wang, Q. Qian, and H. Li, "Instant-teaching: An end-to-end semi-supervised object detection framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4081–4090.

[43] 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.

[44] L. Rossi, A. Karimi, and A. Prati, "Improving localization for semi-supervised object detection," in *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*. Springer, 2022, pp. 516–527.

[45] B. Settles, "Active learning literature survey," 2009.

[46] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[47] P. K. Rhee, E. Erdenee, S. D. Kyun, M. U. Ahmed, and S. Jin, "Active and semi-supervised learning for object detection with imperfect data," *Cognitive Systems Research*, vol. 45, pp. 109–123, 2017.

[48] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[49] 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.

[50] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

[51] 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.

[52] L. Wan, K. Tang, M. Li, Y. Zhong, and A. K. Qin, "Collaborative active and semisupervised learning for hyperspectral remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2384–2396, 2014.

[53] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.