



Addressing Sample Inconsistency for Semisupervised Object Detection in Remote Sensing Images

Yuhao Wang , *Student Member, IEEE*, Lifan Yao , Gang Meng , Xinye Zhang , Jiayun Song ,
and Haopeng Zhang , *Member, IEEE*

Abstract—The emergence of semisupervised object detection (SSOD) techniques has greatly enhanced object detection performance. SSOD leverages a limited amount of labeled data along with a large quantity of unlabeled data. However, there exists a problem of sample inconsistency in remote sensing images, which manifests in two ways. First, remote sensing images are diverse and complex. Conventional random initialization methods for labeled data are insufficient for training teacher networks to generate high-quality pseudolabels. Finally, remote sensing images typically exhibit a long-tailed distribution, where some categories have a significant number of instances, while others have very few. This distribution poses significant challenges during model training. In this article, we propose the utilization of SSOD networks for remote sensing images characterized by a long-tailed distribution. To address the issue of sample inconsistency between labeled and unlabeled data, we employ a labeled data iterative selection strategy based on the active learning approach. We iteratively filter out high-value samples through the designed selection criteria. The selected samples are labeled and used as data for supervised training. This method filters out valuable labeled data, thereby improving the quality of pseudolabels. Inspired by transfer learning, we decouple the model training into the training of the backbone and the detector. We tackle the problem of sample inconsistency in long-tail distribution data by training the detector using balanced data across categories. Our approach exhibits an approximate 1% improvement over the current state-of-the-art models on both the DOTAv1.0 and DIOR datasets.

Index Terms—Active learning, long-tailed distribution, remote sensing, semisupervised object detection (SSOD).

I. INTRODUCTION

Object detection in remote sensing images has undergone substantial advancements in recent years. Numerous remote sensing object detection datasets and detection methods have been constructed and studied, resulting in notable

Manuscript received 11 December 2023; revised 2 February 2024; accepted 3 March 2024. Date of publication 8 March 2024; date of current version 22 March 2024. This work was supported by the National Natural Science Foundation of China under Grant 62271018. (*Corresponding author: Haopeng Zhang.*)

Yuhao Wang is with the Department of Aerospace Information Engineering (Image Processing Center), School of Astronautics, Beihang University, Beijing 102206, China (e-mail: yuhaowang@buaa.edu.cn).

Lifan Yao, Xinye Zhang, and Jiayun Song are with the Qingdao Research Institute of Beihang University, Shandong 266104, China.

Gang Meng is with the Beijing Institute of Remote Sensing Information, Beijing 100011, China.

Haopeng Zhang is with the Department of Aerospace Information Engineering (Image Processing Center), School of Astronautics, Beihang University, Beijing 102206, China, and also with Tianmushan Laboratory, Hangzhou 311115, China (e-mail: zhanghaopeng@buaa.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3374820

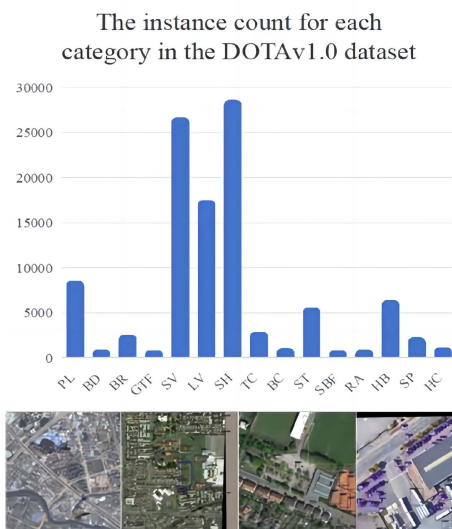


Fig. 1. In remote sensing images, there exists a significant imbalance in the number of instances across different classes. Different images contain varying information and contribute differently to training the network.

achievements. However, the process of instance-level labeling is resource-intensive, posing a hindrance to the enhancement of existing detection models. To address this challenge, semisupervised methods have emerged as a potential solution. These methods enable learning with a limited number of labeled samples and a large pool of unlabeled samples.

Inspired by the successful application of semisupervised learning (SSL) in image classification [1], researchers have extended the teacher–student learning framework to semisupervised object detection (SSOD) and achieved promising outcomes [2], [3]. These algorithms employ a teacher network to generate high-quality pseudolabels. These pseudolabels serve as supervision for training the student network. This self-training approach proves beneficial in scenarios with limited labeled data and large amounts of unlabeled data. Tarvainen et al. [4] introduced the exponential moving average (EMA) technique to the teacher network to alleviate the class imbalance and overfitting problems. Active teacher [5] proposed a dataset-division method based on active learning to enhance the utilization of annotation information.

Although these methods have demonstrated remarkable achievements in the domain of SSOD, they still have certain

limitations for remote sensing images. The issue of sample inconsistency significantly affects the efficacy of semisupervised methods on remote sensing images, as shown in Fig. 1. First, there exists inconsistency in the value of the samples. Different remote sensing images have varying information entropy, object quantity, and categories, resulting in different contributions to training the network. Traditional approaches relying on random initialization of labeled data fail to ensure the maximum contributions. However, the acquisition of knowledge from valuable samples plays a pivotal role in generating high-quality pseudolabels.

Second, the long-tail distribution in remote sensing images is also a manifestation of the sample inconsistency issue. There is a notable disparity in the number of instances across different classes in remote sensing images. Imbalanced categories lead the network to learn a large amount of biased information.

To tackle the inconsistency in sample values, we adopt an iterative approach that focuses on selecting samples with the highest contribution to network learning. This idea originates from the concept of active learning. We extend the traditional teacher–student framework [6] into an iterative framework. The labeled dataset is partially initialized, enhanced, and updated through an active sampling (AS) strategy. This enables the network to maximize the utilization of limited label information and improve the quality of pseudolabels. We employ a comprehensive set of metrics to select samples that offer the most value for SSOD. For the long-tail distribution problem, we use a transfer-learning-based method to address the long-tail problem. High-quality representations can be learned using long-tail datasets. At the same time, we can obtain strong long-tail recognition capabilities by tuning classifiers. Features learned from head classes with abundant training instances can be transferred to under-represented tail classes. We employ a two-stage training strategy to decouple the learning process into a representation learning stage and a detector learning stage. It should be noticed that this article is an extension and improvement of our prior work [7] presented in IGARSS 2023.

In summary, the main contributions of this article are as follows.

- 1) This article introduces an SSOD network specifically designed for remote sensing images, with a focus on addressing the challenge of sample inconsistency. The network tackles the problem of maximizing the value of samples by incorporating an active data selection strategy. Furthermore, it mitigates the impact of the long-tail distribution issue through the utilization of a decoupled training approach.
- 2) The active selection strategy filters out the most valuable samples for annotation based on predefined metrics. The data initialization approach improves the quality of pseudolabels. A two-stage training method is adopted to decouple the learning process into the representation learning stage and the detection learning stage.
- 3) The experimental results prove the state-of-the-art performance of the proposed method. This article also provides a new perspective on the application of SSOD in remote sensing images.

The rest of this article is organized as follows. Section II elucidates the relevant literature on object detection, SSOD, active learning, and long-tail training. Section III presents the methodology and procedure for active data labeling. It also expounds on the training strategies employed to address the long-tail problem. Section IV substantiates the effectiveness of the proposed methods through extensive experimentation. Finally, Section VI concludes this article.

II. RELATED WORK

A. Object Detection

The rapid advancement of deep neural networks has led to remarkable progress in the field of object detection, both in academic research and industrial applications [8], [9], [10], [11], [12], [13]. Object detection methods can be broadly categorized into two genres: one-stage and two-stage detectors. One-stage approaches, such as YOLO [13] and SSD [12], directly predict the object's coordinates and probability distribution from the feature map. On the other hand, two-stage models, exemplified by the RCNN series [9], [11] and its variants [14], employ region proposal networks [15] to sample potential objects, followed by predicting the probability distribution and coordinate information of the objects. In line with prior research efforts [6], [16], our focus lies in the SSL of two-stage models, and we utilize Faster-RCNN [15] as our baseline network.

B. Semisupervised Object Detection

In the domain of computer vision, the predominant focus of current research on SSL is on image classification [17], [18]. This line of inquiry can be broadly divided into two categories: consistency-based and pseudolabeling-based approaches. Consistency-based methods [19], [20], [21], [22] aim to enhance the model's resilience to noise by producing consistent prediction outcomes. Pseudo-labeling methods [16], [17], [22], on the other hand, involve training classifiers with ground-truth annotations and generating pseudolabels for unlabeled data, before ultimately retraining models using all available data.

In recent times, a growing number of research efforts [6], [16], [21], [23] have explored the application of SSL to object detection. For instance, CSD [21] employs multiple random image flips to drive the model toward producing consistent predictions for such flipped images. STAC [16] proposes the first teacher–student-based framework for SSOD. However, the static annotation strategy employed by STAC results in fixed pseudolabels that restrict the final detection performance.

Nevertheless, despite ongoing efforts, the issue of profound instability during the initial training phase persists, necessitating the implementation of a stringent confidence score threshold for the generation of pseudolabels. In an attempt to address these challenges, unbiased teacher [6] leverages the EMA technique [4] to progressively optimize the teacher model based on the knowledge obtained from the student model. Moreover, unbiased teacher also employs EMA [4] in combination with focal loss [24] to effectively tackle the problem of pseudolabel overfitting in the teacher–student learning process.

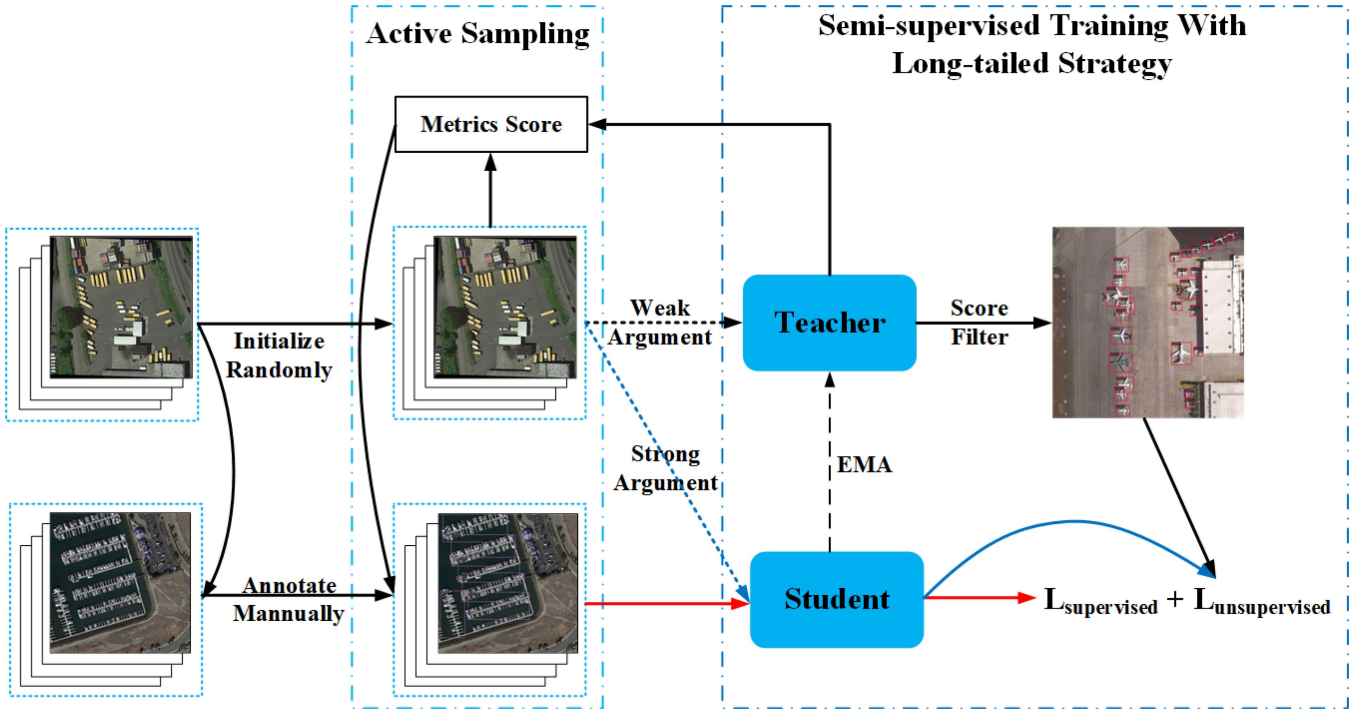


Fig. 2. Overall framework of the proposed model.

C. Active Learning

Efforts toward enhancing learning efficiency have garnered significant attention within the academic community [25], [26]. Moreover, to mitigate the labeling costs associated with object detection, several active-learning-based approaches have been introduced [27], [28], [29]. For example, Wang et al. [27] adopted distinct AS metrics tailored for various stages in the object detection process. CALD [28] assesses information by evaluating the data consistency of bounding boxes before and after augmentation. In addition, MI-AOD [29] employs multi-instance learning techniques to suppress pseudolabel noises effectively.

D. Long-Tailed Training

One line of research aimed at addressing the imbalanced training data is the knowledge transfer from head to tail classes, which has been explored in the literature [30], [31]. Transfer-learning approaches in this context seek to leverage the features learned from head classes, which typically have ample training instances, to benefit the underrepresented tail classes. Recent endeavors in this area encompass methods, such as transferring the intra-class variance [32] and transferring semantic deep features [31]. Nonetheless, it is worth noting that devising a dedicated model for effective feature transfer is often a challenging and intricate task.

In recent studies, it has been demonstrated that the data distribution does not exert any influence on the representation learning of networks [33]. Consequently, there has been a shift toward decoupling the representation and classifier learning processes, leading to notable performance enhancements on

long-tailed datasets. In our model, we also adopt and incorporate this fundamental concept to further improve its efficacy.

III. METHOD

A. Semisupervised Object Detection

The overall framework of our network is illustrated in Fig. 2. The methodology comprises an iterative teacher–student framework and a two-stage training strategy. The limited label set is partially initialized and progressively augmented. Following each iteration, the expert teacher network assesses the significance of unlabeled instances utilizing the proposed metrics, namely, information, diversity, and difficulty. Active data augmentation is subsequently executed based on the evaluation outcomes. For solving the problem of category imbalance, a two-stage training strategy is employed to decouple the learning process into a representation learning stage of the distribution data and a detector learning stage.

Provided a collection of annotated data $D_L = \{X_L, Y_L\}$, alongside a collection of unlabeled data $D_U = \{X_U\}$, where X refers to the examples and Y represents the label set, the objective of SSL is to optimize the performance of the model by leveraging both labeled and unlabeled data.

The proposed SSOD approach incorporates a pair of detection networks, teacher and student network, which share identical configurations, as illustrated in Fig. 2. For the baseline detection network, Faster-RCNN [15] is adopted. Specifically, the teacher network is responsible for generating pseudolabels, whereas the student network is optimized using both ground-truth and pseudolabels.

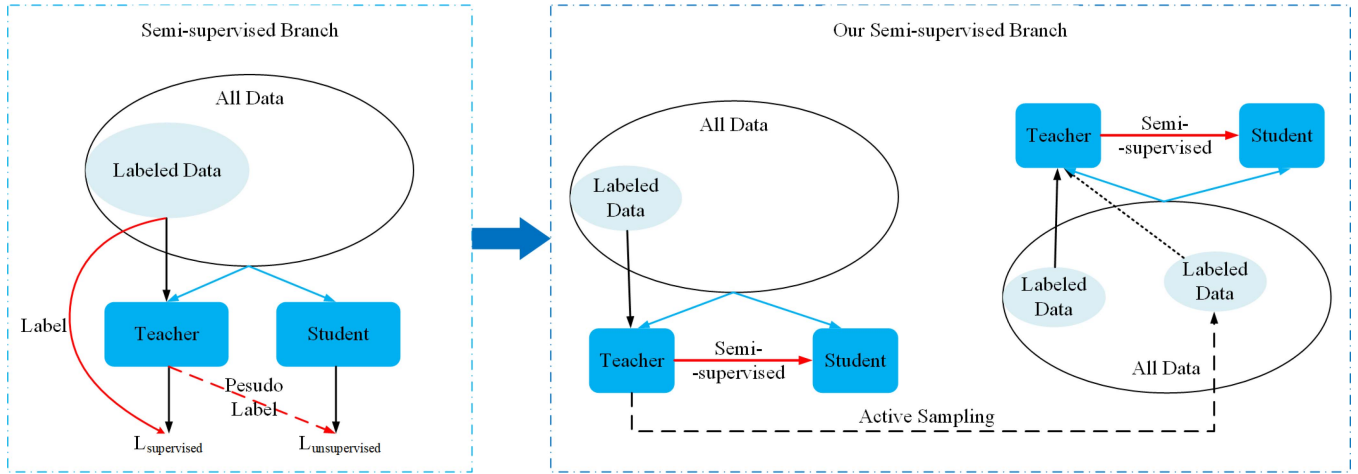


Fig. 3. Comparison of two semisupervised networks. The left network corresponds to the ordinary semisupervised approach, while the right one corresponds to our semisupervised approach.

The optimization loss for the student network can be defined as

$$L = L_{\text{sup}} + \lambda \cdot L_{\text{unsup}} \quad (1)$$

where L_{sup} and L_{unsup} represent the losses corresponding to supervised and unsupervised learning, respectively. λ is the hyperparameter used to balance between L_{sup} and L_{unsup} .

For object detection, L_{sup} comprises the classification loss L_{cls} of RPN and ROI head, in addition to the loss associated with bounding box regression, denoted as L_{loc} . Then, L_{sup} is defined as

$$L_{\text{sup}} = \frac{1}{N_l} \sum_{i=1}^{N_l} (L_{\text{cls}}(x_l^i, y_{\text{cls}}^i) + L_{\text{loc}}(x_l^i, y_{\text{loc}}^i)) \quad (2)$$

where x_l refers to the labeled example, y_{cls} and y_{loc} represent its corresponding classification and bounding box regression labels, respectively. The variable N_l denotes the number of x_l . t_c represents the c th coordinate of the output image x_i . In terms of L_{loc} , the smooth L_1 loss is utilized for bounding box regression, i.e.,

$$\text{Smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{otherwise.} \end{cases} \quad (3)$$

For L_{unsup} , we compute the category loss by comparing the predicted results of the student network with the pseudolabels generated by the RPN and ROI head of the teacher network. It is formulated as

$$L_{\text{unsup}} = \frac{1}{N_u} \sum_{i=1}^{N_u} L_{\text{cls}}(x_u^i, \hat{y}_{\text{cls}}^i) \quad (4)$$

where L_{cls} is the same as (2), and \hat{y}_{cls}^i is the pseudolabels generated by the teacher network.

In order to circumvent the problems of overfitting, a technique proposed in references [6], [23] is employed. The optimization of the teacher network is frozen during semisupervised training,

and its parameters are updated using EMA via

$$\theta_t^i \leftarrow \alpha \theta_t^{i-1} + (1 - \alpha) \theta_s^i \quad (5)$$

where θ_t and θ_s represent the parameters of the teacher and student networks, respectively, and i denotes the i th training step. α is the hyperparameter to determine the speed of parameter transmission, which is normally close to 1. To enhance the quality of pseudolabels, we utilize nonmaximum suppression (NMS) and a confidence threshold to eliminate redundant and ambiguous pseudolabels.

B. Active Sampling

Within the proposed framework shown in Fig. 3, the label set is partially initialized and augmented through the use of the teacher network after each round of semisupervised training. We investigate which examples are essential for SSOD. Three AS metrics, including difficulty, information, and diversity, are employed.

In SSOD, the difficulty score s_i^{diff} of an unlabeled example is calculated based on the category prediction of the teacher network. This score is defined as

$$s_i^{\text{diff}} = -\frac{1}{n_b^i} \sum_{j=1}^{n_b^i} \sum_{k=1}^{N_c} p(c_k; b_j, \theta_t) \log p(c_k; b_j, \theta_t) \quad (6)$$

where n_b^i represents the number of predicted bounding boxes after NMS and confidence filtering. N_c denotes number of object categories, and $p(c_k; b_j, \theta_t)$ is the prediction probability of the k -th category by the teacher network. Referencing (6), the prediction uncertainty of the teacher network can be utilized to determine whether an image presents a challenge for SSOD.

Information serves as a metric used to quantify the amount of information contained in an unlabeled image for SSOD. In the context of object detection, a higher level of information richness indicates the presence of an increased number of visual concepts within the image. It enables the model to learn more

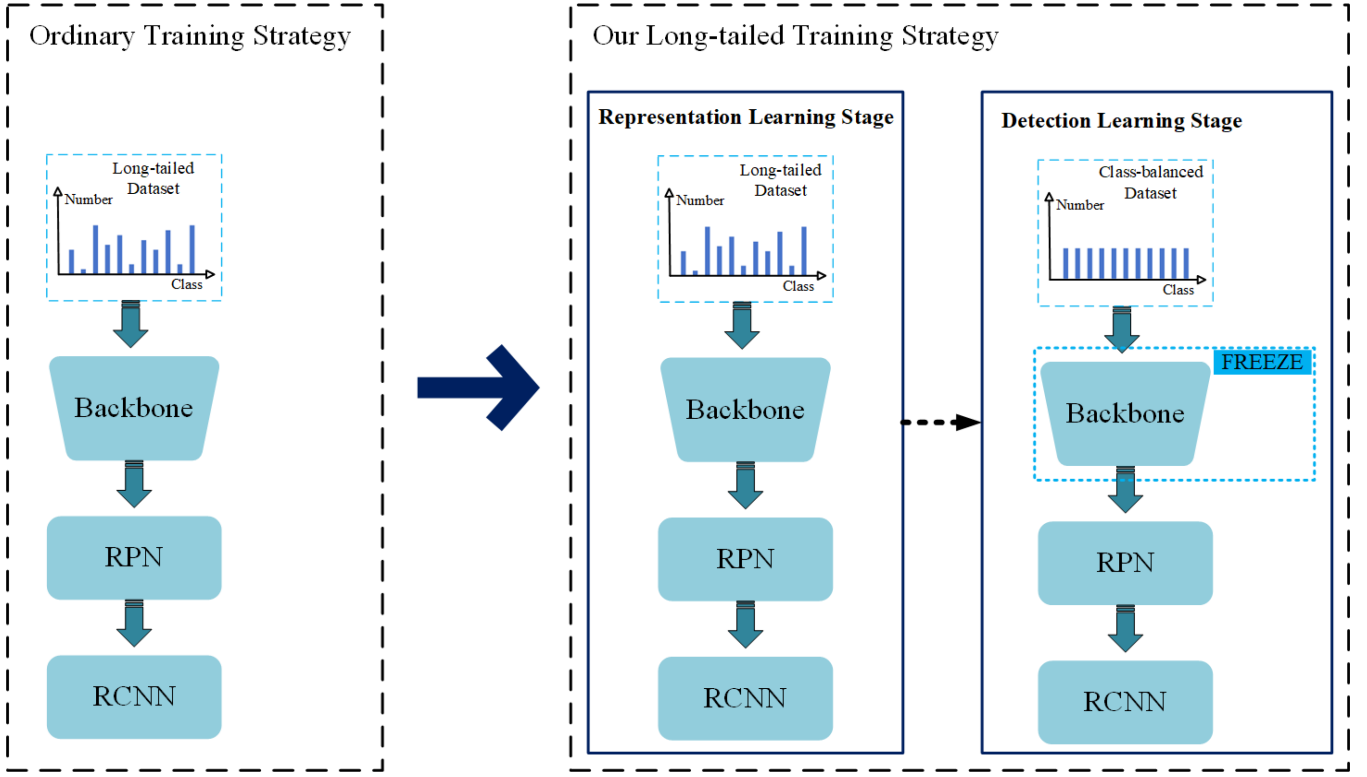


Fig. 4. Framework of the LTS.

features for object detection. Thus, we utilize prediction confidence to evaluate this metric as follows:

$$s_i^{\text{info}} = \sum_{j=1}^{n_b^i} \text{confidence}(b_j, \theta) \quad (7)$$

where the $\text{confidence}(b_j, \theta_t)$ represents the highest confidence score among the predicted bounding boxes by the teacher network.

Diversity serves as a metric for evaluating the distribution of object categories within an image. The diversity score s_i^{dive} is computed by

$$s_i^{\text{dive}} = \left| \{c_j\}_{j=1}^{n_b^i} \right| \quad (8)$$

where c_j represents the predicted category of the j th bounding box, and $|\cdot|$ denotes the cardinality.

The proposed metrics have the potential to address the issue of determining suitable examples for SSOD. However, a practical challenge emerges due to the fact that models in different states may require varying levels of label information. In addition, optimizing the benefits of these metrics without conducting extensive experimentation presents a significant hurdle. Hence, we introduce a simple yet effective method to automatically integrate these metrics.

Prior to the combination, the values are normalized as

$$s_i^{\hat{m}} = \frac{s_i^m}{s_{\max}^m} \quad (9)$$

where $m \in \{\text{difficulty, information, diversity}\}$ represents the metrics. s_{\max}^m denotes the maximum value of this metric.

Considering that these metrics capture image information from different perspectives, we proceed to construct a 3-D sampling space to represent each example, represented by $\vec{s}_i = \{s_i^{\text{diff}}, s_i^{\text{info}}, s_i^{\text{dive}}\}$. The evaluation outcome of each unlabeled example can be viewed as a point in this space. Then, we apply L_p normalization to the data points, resulting in a single scalar value s_{L_p} , i.e.,

$$s_{L_p} = l_p(\hat{s}) = \|\hat{s}\|_p = \sqrt[p]{\sum_{i=1}^3 s_i^p} \quad (10)$$

where $\hat{s} = (s_i^{\text{diff}}, s_i^{\text{info}}, s_i^{\text{dive}})$. Empirically, we utilize the L_1 norm to combine these three metrics.

C. Long-Tailed Training Strategy

In the remote sensing images, the issue of long-tail distribution presents a significant challenge. This phenomenon refers to the occurrence where a large number of classes in the dataset are represented by a relatively small number of instances, while a few classes are represented by a large number of instances. This skewed distribution leads to a scenario where the majority of classes have insufficient examples, complicating the process of accurate classification and analysis. The long-tail distribution in remote sensing images hampers the efficacy of SSOD algorithms, which typically presume a relatively balanced dataset. It necessitates the development of

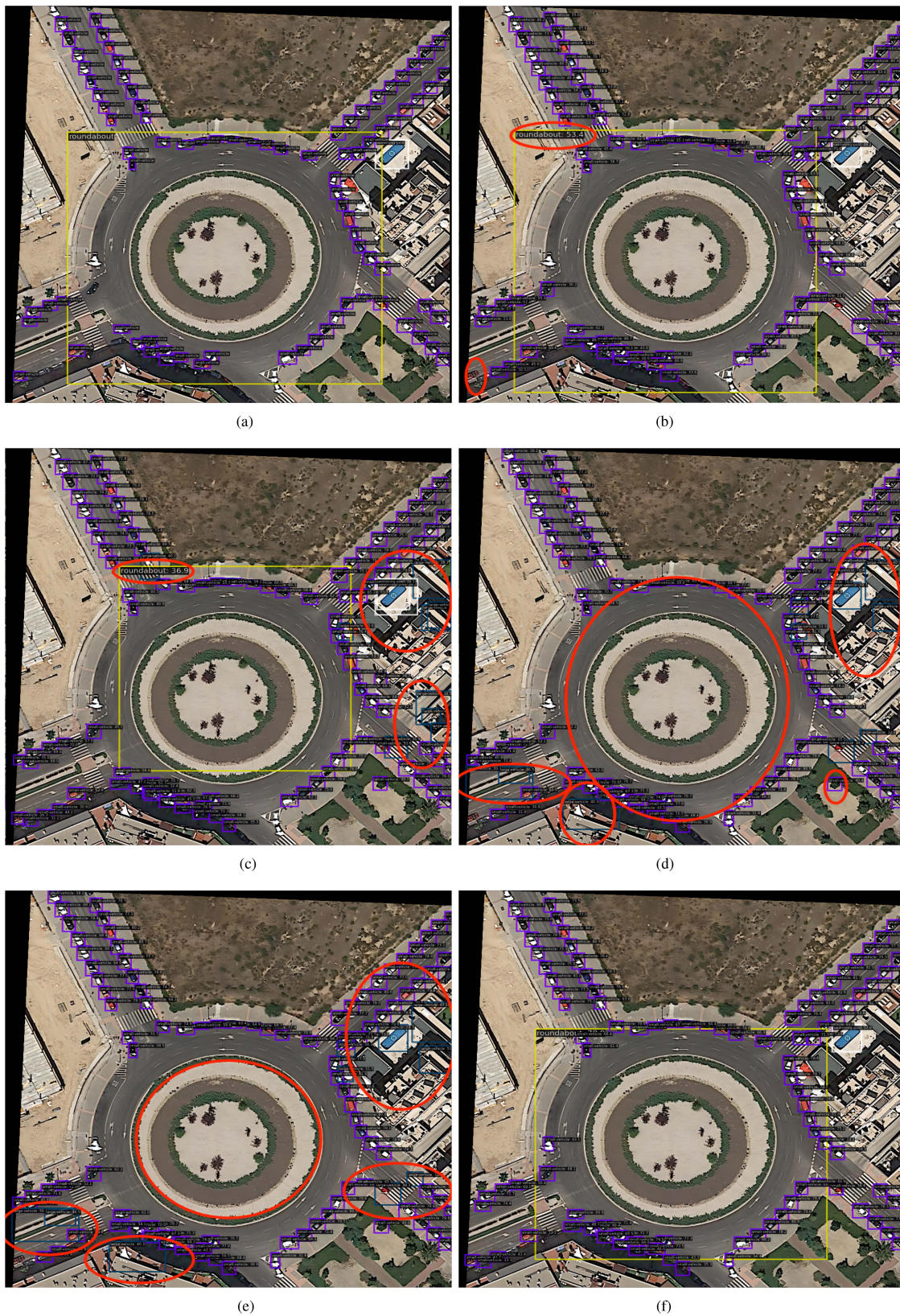


Fig. 5. Experimental results on DOTAv1.0. Experiments were conducted with a setting of 20% labeled data. The red circles, respectively, indicate the missed objects, falsely detected objects, and low-scoring objects within the image. (a) Ground Truth. (b) EPC. (c) Active Teacher. (d) Soft Teacher. (e) Unbiased Teacher. (f) Ours.

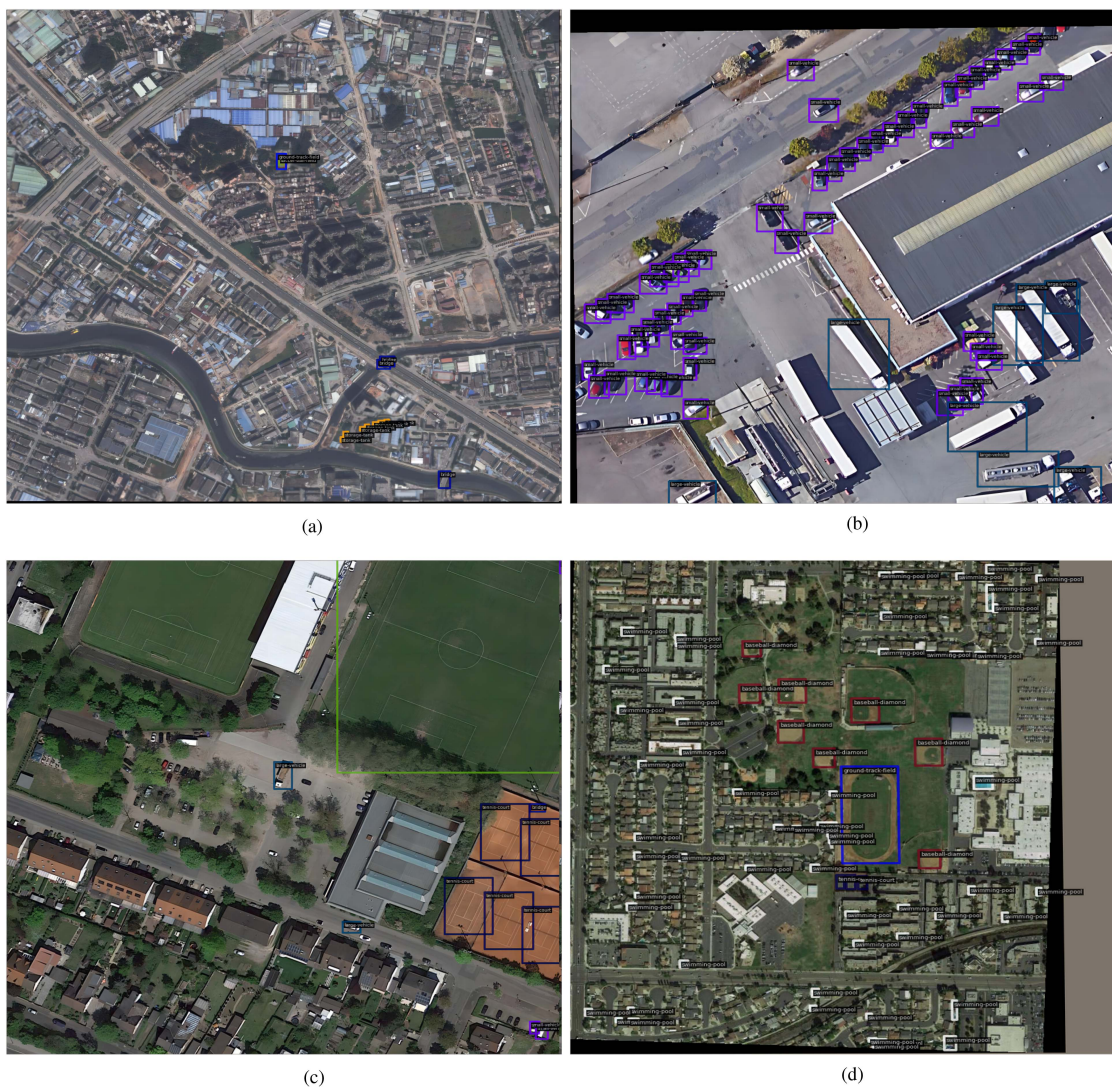


Fig. 6. Samples selected by different metrics. (a) Sample selected by difficulty. (b) Sample selected by information. (c) Sample selected by diversity. (d) Sample selected by combination metrics.

specialized approaches that can effectively handle the disparity in sample sizes across various classes, thereby ensuring a more robust and comprehensive analysis of the remote sensing data.

For the class imbalance problem, previous work [34] shows that high-quality representations can be learned using long-tail datasets, and at the same time, strong long-tail recognition capabilities can be obtained by tuning classifiers using only class-balanced datasets. In remote sensing images, the issue of long-tail distribution does not hinder the learning of features. The network can utilize all the data for thorough feature extraction and learning. With a substantial amount of data, the network is capable of learning both low-level and high-level features of remote sensing objects. However, in specific classification and detection processes, the long-tail distribution has a significant impact on the detector. Therefore, it is necessary to create a specially balanced subdataset to train the detector separately. To retain the vast amount of effective and diverse

features previously learned by the network, fine-tuning on a pretrained model presents an ideal solution. This approach not only maximizes the preservation of knowledge previously acquired through extensive data but also mitigates the limitations imposed by the long-tail distribution problem on the detector.

Inspired by this, we employ a two-stage training strategy to decouple the learning process into a representation learning stage of the distribution data and a detector learning stage shown as Fig. 4. In the representation learning phase, neural networks are trained on a given dataset. During the detector learning phase, only the detector parameters are updated for retraining by freezing the backbone parameters on the class-balanced dataset. We construct these class-balanced datasets from given labeled data without additional images.

Utilizing data augmentation during the testing phase yields improved overall performance. Nonetheless, owing to the influence of the long-tail constraint, the model inclines toward

TABLE I
EXPERIMENTAL RESULTS ON DOTAV1.0. THE EVALUATION METRIC IS MAP

	10% of samples	20% of samples	40% of samples
Supervised	50.21	61.26	67.25
Unbiased teacher [6]	57.85(+7.64)	67.34(+5.88)	71(+3.75)
Soft teacher [40]	62.45(+12.24)	70.78(+9.32)	73.96(+6.71)
Active teacher [5]	64.2(+13.99)	72.3(+11.04)	76.5(+9.25)
EPC [41]	64.7(+14.49)	72.93(+11.67)	76.45(+9.2)
Ours	65.34(+15.13)	73.68(+12.22)	76.87(+9.62)

Bold fonts mean the best performance.

TABLE II
EXPERIMENTAL RESULTS ON DIOR. THE EVALUATION METRIC IS MAP

	10% of samples	20% of samples	40% of samples
Supervised	49.35	58.99	66.73
Unbiased teacher [6]	54.75(+6.40)	64.77(+5.78)	69.35(+2.62)
Soft teacher [40]	59.63(+10.28)	66.03(+7.04)	72.30(+5.57)
Active teacher [5]	62.1(+12.75)	68.4(+9.41)	73.72(+6.99)
EPC [41]	63.3(+13.95)	68.98(+9.99)	73.02(+6.29)
Ours	62.47(+13.12)	69.76(+10.77)	73.94(+7.21)

Bold fonts mean the best performance.

TABLE III
EFFECTS OF AS AND LTS ARE STUDIED IN EXPERIMENTS CONDUCTED ON
10% DOTAV1.0 LABELED DATA SETTINGS

AS	LTS	mAP
		57.85
✓		64.2
	✓	63.05
✓	✓	65.34

Bold fonts mean the best performance.

predicting samples as belonging to the dominant class. Consequently, the efficacy of test-time augmentation (TTA) is diminished. To address this concern, we employ a nonparametric postprocessing methodology known as classification with alternating normalization (CAN [35]).

IV. EXPERIMENT

A. Datasets and Evaluation Metrics

In order to demonstrate the practicality and generalization of the model proposed in this article, experiments were conducted on two datasets, namely, the DIOR dataset [36] and the DOTAv1.0 [37] dataset.

1) *DOTAv1.0*: [37] contains 2806 large-scale aerial images with sizes ranging from 800×800 to 4000×4000 . It consists of 188 282 instances, including airplanes (PL), baseball diamonds (BD), bridges (BR), ground track fields (GTF), small vehicles (SV), large vehicles (LV), ships (SH), tennis courts (TC), basketball courts (BC), storage tanks (ST), soccer fields (SBF), roundabouts (RA), harbors (HA), swimming pools (SP), and helicopters (HC).

2) *DIOR* (See [36]): Serves as a substantial and openly accessible benchmark for object detection tasks in optical remote sensing images. It comprises 23 463 images and a total of 192 472 instances, covering a diverse range of 20 distinct object classes.

During our experimental process, the split is further divided into two subsets: the labeled set and the unlabeled set. This division follows a similar approach to prior studies in SSOD. In our practical implementation, we employ various proportions of labeled data from the DOTAv1.0 and DIOR dataset, namely 10%, 20%, and 40%, for conducting experiments and comparing the results with other SSOD methods. The remaining examples are considered as unlabeled data.

For model evaluation, we adhere to the methodologies employed in previous studies [6], [16], [21], [23] and utilize mean average precision (mAP) with intersection over union (IoU) thresholds ranging from 0.5 to 0.95 as the evaluation metric for our experiments. In addition, the validation sets of DOTAv1.0 and DIOR are employed for evaluation purposes.

B. Implementation Details

In accordance with prior studies [38], [39], a cropping technique is employed to divide the original images into patches of size 1024×1024 , with a stride of 824 pixels. It results in a pixel overlap of 200 between adjacent patches. In order to augment the unlabeled data, we utilize an asymmetric data augmentation approach. All models undergo training for a total of 180 000 iterations using 2 GTX2080 GPUs. The stochastic gradient descent (SGD) optimizer is employed with an initial learning rate of 0.001, which is reduced by a factor of 10 at iterations 120 000 and 160 000. The momentum is set to 0.9, while the weight decay is set to 0.0001. The EMA is configured with $\alpha = 0.9996$, and the unsupervised loss is assigned a weight parameter of $\lambda = 4$ for all experiments. In all conducted experiments, half of the labeled dataset is randomly selected, while the remaining half is actively chosen after the process of SSL. The batch size is set to eight, comprising four labeled images and four unlabeled images selected through random sampling.

C. Results on DOTAv1.0

Initially, we conducted a comparison between our method and a range of teacher–student based SSOD methods on DOTAv1.0, the outcomes of which are presented in Table I. For a fair comparison, we reimplemented these methods with the same settings. We evaluated our method under different labeled data proportions. The aforementioned table indicates that all the teacher–student based techniques significantly outperform the conventional supervised learning approach. Furthermore, we observed that the aforementioned teacher–student methods, such as active teacher [5], which have been recently proposed, exhibited marked improvements over the pioneering method, namely, unbiased teacher [6], owing to their meticulous design frameworks. This observation underscores the noteworthy advancements in teacher–student based SSOD. Our method achieves state-of-the-art performance under all proportions.

The method achieves good performance under all proportions. Specifically, it obtains 65.34%, 73.68%, and 76.87% mAP on 10%, 20%, and 40% proportions, respectively, surpassing our supervised baseline by +15.13, +12.22, and +9.62 mAP. Furthermore, we outperform the method EPC [41] by +0.64, +0.75, and +0.42 under various proportions.

TABLE IV
ABLATION EXPERIMENTS OF LTS ON DOTAV1.0; EXPERIMENTS ARE PERFORMED USING 10% OF LABELED DATA

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Without LTS	85.5	68.4	47.9	56.1	59.4	83.2	75.6	94.1	62.8	71.8	58.3	54	75.7	53.2	16.7	64.2
With LTS	86	69.2	49.1	57.4	61.1	84.5	76.7	95.2	63.2	73.4	59.5	55.6	76.5	54.1	18.6	65.34



Fig. 7. Experimental results on DIOR were conducted using a 20% labeled data setting. Each line represents the ground truth on the left side, while the predicted visualization results are shown on the right side.

TABLE V
EFFECTS OF AS ARE INVESTIGATED IN EXPERIMENTS CONDUCTED ON 10% DOTAV1.0 LABELED DATA SETTINGS

Method	mAP
With difficulty metrics	63.75
With information metrics	63.67
With diversity metrics	63.15
With L1 combination metrics	65.34
With L2 combination metrics	64.25

Bold fonts mean the best performance.

The qualitative results of our method compared with other networks are shown in Fig. 5. The proposed model in this article can accurately detect large-scale roundabout targets and dense small-scale small vehicle targets, while also detecting swimming pool targets in complex backgrounds. In addition to achieving high accuracy in category recognition, the regression of detection boxes is also highly precise. With the help of our AS, the model is able to exploit more potential semantic information from the unlabeled data, helping reduce false predictions and improve the pseudolabel quality. By employing the long-tailed training strategy, the model can effectively alleviate the problem of category imbalance. Furthermore, our proposed model can identify missing annotated instances, which demonstrates that the model has good generalization and robustness.

D. Results on DIOR

Furthermore, we performed a comparative analysis between our approach and the SSOD methods on the DIOR dataset. To ensure a fair and unbiased comparison, we reimplemented these methods on object detectors using identical augmentation settings. We evaluated the performance of our method across different proportions of labeled data. The corresponding results are presented in Table II.

Our method achieved mAP values of 62.47%, 69.76%, and 73.94% when evaluated on labeled data proportions of 10%, 20%, and 40%, respectively. These results demonstrate an improvement of +13.12, +10.77, and +7.21 mAP over our supervised baseline. In addition, our approach outperformed the active teacher [5] by +0.37, +1.36, and +0.22 for the different data proportions.

Fig. 7 shows that the proposed method achieves precise bounding box regression and accurate class recognition for detected objects. Specifically, the network is able to accurately detect objects such as aircraft targets of various sizes and densely packed small targets like vehicles. This is attributable to the AS module, which selects labels that provide the maximum training assistance. It enhances feature extraction and semantic learning capabilities and improves the quality of label generation. In addition, the LTS module helps in balancing the detection biases among different object categories. Overall, our proposed method has achieved superior results compared to the baseline approaches on two well-known remote sensing datasets. These results effectively demonstrate the effectiveness, robustness, and practicality of our method.

E. Ablation Study

In this section, we conduct extensive studies to validate each of our each modules. Unless specified, all ablation experiments are performed using 10% of labeled data.

1) *Ablation Study of Each Component*: We have conducted a study to analyze the effects of the proposed methods, AS and LTS. As shown in Table III, both methods have been proven to be effective and complementary. AS and LTS individually contribute to performance gain. When combined, AS and LTS synergistically enhance the performance of the baseline model. It indicates that the active labeled data sampling by AS and the decoupled training of the backbone and detector by LTS can benefit the SSL process. These methods facilitate the generation of high-quality pseudolabels and mitigate the impact of class imbalance on the detector.

Fig. 8 illustrates that the incorporation of active learning enables the selection of images that cover a comprehensive range of difficulties, information content, and class quantities. This, in turn, enhances the quality of the pseudolabels used for

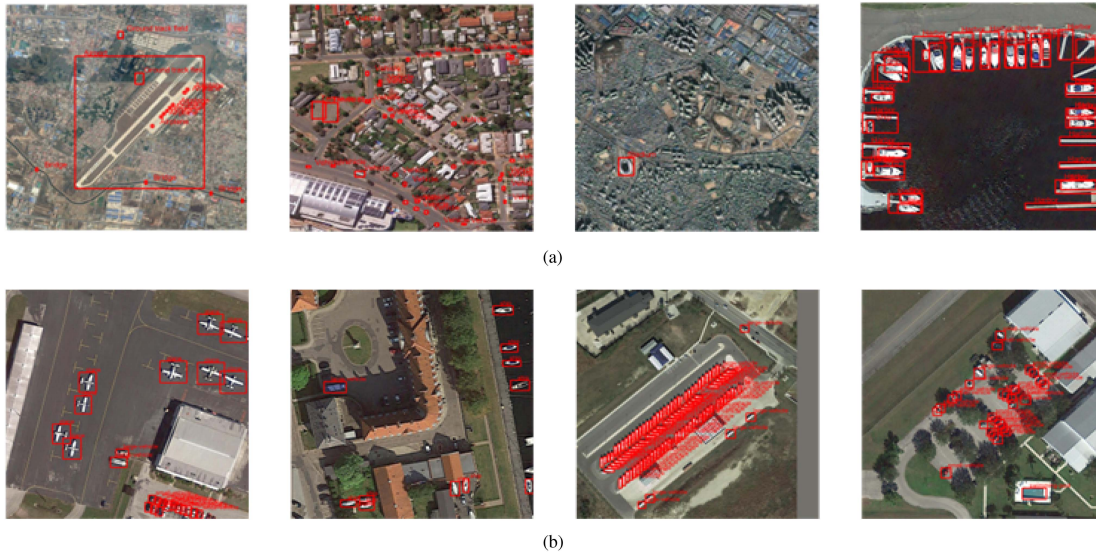


Fig. 8. Samples selected by active learning. (a) Selected Samples of DOTAv1.0. (b) Selected Samples of DIOR.

model training. As a result, the model demonstrates exceptional performance in detecting dense objects, complex background objects, and small objects.

In Table IV, we observe that the inclusion of the LTS module improves the model's mAP to 65.34, which is a 1.14 improvement. Notably, classes with limited instances, such as baseball-diamond, ground-track-field, soccer-ball-field, roundabout, and helicopter, experience an average improvement of 1.2. The decoupling training approach, based on transfer learning, effectively enhances the learning of rich image features and semantic information while mitigating the bias caused by imbalanced instance quantities during the training process. As a result, the approach significantly improves detection accuracy.

2) *Ablation Study of AS*: Regarding the selection metrics in active learning strategies, this article conducted detailed ablation experiments. The results from the Table V demonstrate that using the individual metrics of diff, info, and div for image selection yielded training results of 63.75, 63.67, and 63.15, respectively. However, the joint selection strategy combining all three metrics shows an average improvement of 2% points on mAP. Furthermore, the strategy used L1 norm outperforms the strategy used L2 norm.

As shown in Fig. 6, the difficulty metric primarily filters images in which instances are difficult to detect, such as those with complex backgrounds and small targets. The information metric filters images with a higher number of instances, often including densely populated objects such as small vehicles or ships. The diversity metric selects images that contain rich category information. Our proposed combined filtering strategy integrates these three metrics to comprehensively filter images, initially selecting a subset of images that can provide the maximum training benefits to the model.

V. DISCUSSION

This article effectively addresses the issue of sample inconsistency in SSOD in remote sensing images. In remote

sensing images, objects often have specific orientations, which are overlooked by the horizontal bounding boxes used in this study. Remote sensing objects are densely arranged and have specific orientations, which poses significant challenges to SSOD. However, the angle information of the objects can also provide richer features and learnable knowledge for SSOD. Considering the characteristics of remote sensing objects, future research could explore how to design an effective semisupervised oriented object detection network.

VI. CONCLUSION

In this article, we propose an SSOD framework for remote sensing data, addressing the issue of sample inconsistency. We present a data initialization method in the student network based on the concept of active learning and conduct extensive experiments to select sampling indices and strategies. Our approach effectively tackles the problem of maximizing the value of samples in the context of sample inconsistency. Furthermore, the issue of sample inconsistency is also manifested in the category imbalance of remote sensing data. To address this issue, the teacher–student semisupervised network adopts a two-stage training strategy, decomposing the learning process into a representation learning stage and a detector learning stage. Experimental results demonstrate the state-of-the-art performance of the proposed method and offer a novel design approach for semisupervised object detection in remote sensing images.

REFERENCES

- [1] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020.
- [2] B. Xue and N. Tong, "Diod: Fast and efficient weakly semi-supervised deep complex ISAR object detection," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3991–4003, Nov. 2019.

- [3] B. G. Weinstein, S. Marconi, S. Bohlman, A. Zare, and E. White, "Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1309.
- [4] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1703–1712.
- [5] P. Mi et al., "Active teacher for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14482–14491.
- [6] Y.-C. Liu et al., "Unbiased teacher for semi-supervised object detection," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 2102–2113.
- [7] Y. Wang, L. Yao, G. Meng, X. Zhang, J. Song, and H. Zhang, "Semi-supervised object detection in remote sensing images based on active learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 5571–5574.
- [8] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOx: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
- [9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
- [11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
- [12] W. Liu et al., "Ssd: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
- [13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
- [14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
- [15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 1137–1149, 2015.
- [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," 2020, *arXiv:2005.04757*.
- [17] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 3365–3373, 2014.
- [18] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 7110–7122.
- [19] D. Berthelot et al., "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 6832–6844.
- [20] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 2314–2322, 2019.
- [21] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 1132–1141, 2019.
- [22] K. Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 596–608, 2020.
- [23] Y. Tang, W. Chen, Y. Luo, and Y. Zhang, "Humble teachers teach better students for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3132–3141.
- [24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
- [25] S. Zhang, F. Jia, C. Wang, and Q. Wu, "Targeted hyperparameter optimization with lexicographic preferences over multiple objectives," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 777–788.
- [26] X. Zheng et al., "DDPNAS : Efficient neural architecture search via dynamic distribution pruning," *Int. J. Comput. Vis.*, vol. 131, no. 5, pp. 1234–1249, 2023.
- [27] Z. Wang, Y. Li, Y. Guo, L. Fang, and S. Wang, "Data-uncertainty guided multi-phase learning for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4568–4577.
- [28] W. Yu, S. Zhu, T. Yang, and C. Chen, "Consistency-based active learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3951–3960.
- [29] T. Yuan et al., "Multiple instance active learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5330–5339.
- [30] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 674–685, 2017.
- [31] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2537–2546.
- [32] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5704–5713.
- [33] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9719–9728.
- [34] B. Kang et al., "Decoupling representation and classifier for long-tailed recognition," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1432–1443.
- [35] M. Jia, A. Reiter, S.-N. Lim, Y. Artzi, and C. Cardie, "When in doubt: Improving classification performance with alternating normalization," *Findings Assoc. Comput. Linguistics*, pp. 1154–1163, 2021.
- [36] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.
- [37] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
- [38] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2021.
- [39] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2786–2795.
- [40] M. Xu et al., "End-to-end semi-supervised object detection with soft teacher," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3060–3069.
- [41] J. Shen, C. Zhang, Y. Yuan, and Q. Wang, "Enhancing prospective consistency for semi-supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, to be published.



Yuhao Wang (Student Member, IEEE) received the B.S. degree in aircraft control and information engineering in 2022 from Beihang University, Beijing, China, where he is currently working toward the master's degree in electronic information engineering.

His main research interests include remote sensing image processing, semisupervised learning, and object detection.



Lifan Yao received the B.S. degree in electronic information engineering from Beihang University, Beijing, China, in 2010.

He studied with the University of Nottingham during the period from 2008 to 2009. He is currently an Executive Dean of Aerospace Technology Application Research Institute, Qingdao Research Institute, Beihang University. His main research interests include remote sensing image processing and analysis, remote sensing data fusion, and applications.

Mr. Yao is a Senior Member of the Chinese Society

of Astronautics.



Gang Meng received the Ph.D. degree in pattern recognition and intelligent system from Beihang University, Beijing, China, in 2011.

He is currently a Senior Engineer with the Beijing Institute of Remote Sensing Information, Beijing. His research interests include intelligent object detection and recognition in remote sensing images, pattern recognition, and machine learning.



Xinye Zhang received the B.S. degree in computer technology from Qingdao University of Science and Technology, Qingdao, China, in 2020.

He is currently an Intermediate Engineer with Qingdao Research Institute, Beihang University, Beijing, China. He mainly engages in the application research of artificial intelligence in the field of aerospace. His main research interests include image processing and computer vision.



Jiayun Song received the B.S. degree in business administration from Beijing International Studies University, Beijing, China, in 2019.

She is currently an Intermediate Engineer with Qingdao Research Institute, Beihang University, Beijing. She mainly engages in artificial intelligence and commercial aerospace industry development research.



Haopeng Zhang (Member, IEEE) received the B.S. degree in detection, guidance and control technology and Ph.D. degree in pattern recognition and intelligent system from Beihang University, Beijing, China, in 2008 and 2014, respectively.

He is currently an Associate Professor with the Department of Aerospace Information Engineering (Image Processing Center), School of Astronautics, Beihang University. His main research interests include remote sensing image processing, multiview object recognition, 3-D object recognition and pose estimation, and other related areas in pattern recognition, computer vision, and machine learning.