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

Boxuan Zhang, Zengmao Wang and Bo Du, *Senior Member, IEEE*

Abstract—The lack of object-level annotations poses a significant challenge for object detection in remote sensing images. To address this issue, active learning and semi-supervised learning techniques have been proposed to enhance the quality and quantity of annotations. Active learning focuses on selecting the most informative samples for annotation, while semi-supervised learning leverages the knowledge from unlabeled samples. In this paper, we propose a novel active learning method to boost semi-supervised object detection for remote sensing images with a teacher-student network, called SSOD-AT. The proposed method incorporates a RoI Comparison module (RoICM) to generate high-confidence pseudo-labels for Regions of Interest (RoIs). Meanwhile, the RoICM is utilized to identify the top-K uncertain images. To reduce redundancy in the top-K uncertain images for human labeling, a diversity criterion is introduced based on object-level prototypes of different categories using both labeled and pseudo-labeled images. Extensive experiments on DOTA and DIOR two popular datasets demonstrate that our proposed method outperforms state-of-the-art methods for object detection in remote sensing images.

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

I. INTRODUCTION

OBJECT detection in remote sensing images is a crucial task to identify the objects with their locations [1]. In recent years, deep learning has shown promising results in object detection for remote sensing images. However, the success of deep learning approaches heavily relies on large-scale datasets with accurately labeled data, which are typically annotated by human experts. Unlike image classification, object detection requires object-level labels that include both bounding box coordinates and object categories. This annotation process is more laborious and time-consuming. Additionally, remote sensing images often contain objects with various orientations and scales [2], [3], [4], [5], [6], which further complicates the labeling process. Hence, although the collection of remote sensing images are to be faster and easier with the advancements of remote sensing technology, the availability of labeled images for object detection is usually 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. For the semi-supervised object detection(SSOD) 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 [10]. The student network then is optimized by using the pseudo-labels as supervised information.

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 [11]. The query criterion is the core technique in AL methods, i.e. uncertainty and diversity [12], [13]. Yuan et.al [14] 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 [15] not only gauges individual information for sample selection but also leverages mutual information to alleviate unbalanced class distribution, thus ensuring the diversity of selected samples.

Therefore, it is natural to consider combining AL and SSL to enhance the performance of object detection when dealing with limited labeled Remote Sensing Images (RSIs). In fact, several methods have been developed for Semi-Supervised Object Detection (SSOD) with AL in natural images. For instance, Lv et al. proposed a novel semi-supervised active salient object detection (SOD) method that utilized a salient encoder-decoder with an adversarial discriminator to select the most representative data. Mi et al. [16] introduced the concept of "Active Teacher," which involves data initialization through AL using a teacher-student-based SSOD approach. However, these methods may not be directly applicable to remote sensing images due to their higher resolution and the presence of more complex objects.

We note that SSL usually relies on the labeled data to exploit the unlabeled data for task learning, while AL aims to label the most informative samples from the unlabeled data with a query criterion. Therefore, it is a natural consideration to combine AL and SSL together to improve the performance of object detection with limited labeled RSIs. In fact, many methods have been developed for SSOD with AL in natural images. Lv et al. proposed a novel semi-supervised active salient object detection (SOD) method that utilized a salient encoder-decoder with an adversarial discriminator to select the most representative data. Mi et al. [16] introduced the method "Active Teacher", which involves data initialization through AL using a teacher-student-based SSOD approach. However, these methods usually ignore the redundancy in the ROIs, resulting in that more images are selected to improve the performance for SSOD.

In this paper, we propose a method Semi-Supervised Object Detection with Active Teaching, termed as SSOD-AT, for object detection in remote sensing images using a teacher-
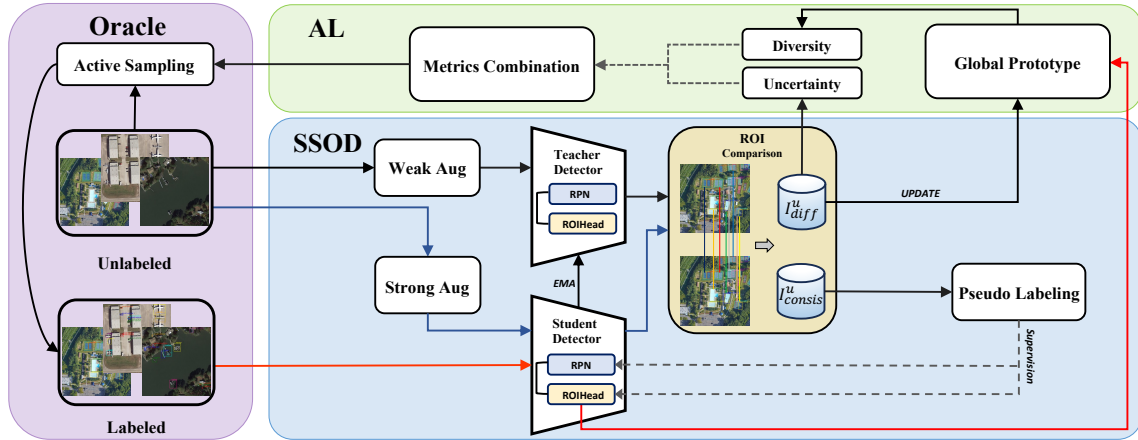
Fig. 1. Overview of SSOD-AT framework for three stages. **Semi-Supervised Object Detection(SSOD)**: Using limited label set to initialize the parameters of Teacher-Student framework. **Active Learning(AL)**: Select the top-N valuable samples for labeling. **Label set Augmentation(Oracle)**: Using the active selected samples to augment the label set. Repeat the preceding procedures to train the Teacher-Student framework.

student network. Remote sensing images often contain objects with large scale variations and high density, which is easy to produce noisy ROIs. To mitigate the impact of these noisy RoIs, we introduce a RoI Comparison Module (RoICM) that compares the RoIs generated by the teacher network and the student network. In our method, when the RoIs exhibit consistent predictions between the teacher and student networks, they are assigned pseudo-labels for semi-supervised training. Conversely, for RoIs with divergent predictions, we utilize them to calculate the uncertainty of the image based on the teacher network's predictions. To remove the redundancy from the queried images, a diversity is designed by introducing a global class prototype to ensure the diverse between the current query image and selected images. Finally, we integrate the uncertainty and diversity as the query score to select the most valuable images for human labeling. The main contributions of this article can be summarized as follows:

- A novel method to boost semi-supervised object detection with active learning is proposed for remote sensing images based on the teacher-student network. The proposed method can provide both confident pseudo-labels and informative images.
- A RoI comparison module(RoICM) is introduced by comparing the RoIs generated by teacher and student network. It can effectively alleviate the influence of noisy RoIs for semi-supervised learning and improve the ability of active learning to evaluates the uncertainty of images.
- The proposed method further incorporates the global class prototype for the diversity of selected images. The combination of the two sampling strategies maximizes the effectiveness of AL process.

II. METHODOLOGY

The proposed SSOD-AT method is shown in Fig.1. We introduce an iterative strategy to train Teacher-Student network with active learning for semi-supervised object detection. In SSOD-At, a RoI Comparison Module is introduced by producing confident pseudo-labels for the ROIs and providing candidate ROIs for active learning to measure the uncertainty

of the images. Meanwhile, a diversity criterion is designed based on the *Global Class Prototype* to remove the redundancy with the queried images. In this section, we will introduce the details of the *RoI Comparison Module(RoICM)* and the diversity in detail.

A. RoI Comparison Module(RoICM) for Uncertainty selection

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: 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. 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.

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:

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

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 $p^i(b_j, c_k)$ 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. A lower

D_{kl} indicates more reliable pseudo-labels generated by the teacher. Thus we assign a higher weight to those of images when calculating the total loss as follows:

$$\mathcal{L}_{det}^{ts}(\mathcal{D}_l, \mathcal{D}_u) = \mathcal{L}_{det}^{sup}(\mathcal{D}_l) + \exp(-D_{kl}) \cdot \lambda_u \cdot \mathcal{L}_{det}^{unsup}(\mathcal{D}_u) \tag{2}$$

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:

$$\cdot S_{unc}^i = -\frac{1}{n_b^i} \sum_{j=1}^{n_b^i} \sum_{k=1}^{N_c} p_t^i(b_j, c_k) \log p_t^i(b_j, c_k) \tag{3}$$

B. Global Class Prototype for Diversity selection

We have established a global prototype for each class, which serves as the basis for ensuring the diversity of the selected image categories. Given a labeled image x_i^l, we obtain a set of RoI features $\mathbf{F}_i^{gt}$ from the ground-truth bounding boxes generated by the RoI Head of the Student Detector in each training stage:

$$\mathbf{F}_i^{gt} = \left\{ \left(f_{i,j}^{gt}, y_{i,j}^{gt} \right) \right\} \tag{4}$$

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:

$$v_k = \begin{cases} \frac{\sum_{i,j} f_{i,j}^{gt} \mathbb{1}\left(y_{i,j}^{gt}=k\right)}{\sum_{i,j} \mathbb{1}\left(y_{i,j}^{gt}=k\right)} & \sum_i \mathbb{1}\left(y_{i,j}^{gt}=k\right) > 0 \\ \mathbf{0} & \sum_i \mathbb{1}\left(y_{i,j}^{gt}=k\right) = 0 \end{cases} \tag{5}$$

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:

$$\mathbb{1}\left(y_{i,j}^{gt}=k\right) = \begin{cases} 1 & \text{if } y_{i,j}^{gt} = k \\ 0 & \text{if } y_{i,j}^{gt} \neq k \end{cases} \tag{6}$$

The global class prototype is updated using the EMA algorithm [17], with the local class prototype serving as a reference:

$$g_k = \alpha g_k + (1 - \alpha) v_k \tag{7}$$

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 class 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 class 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. 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:

$$\text{sim}\left(f_{i,j}^{pgt}, g_k\right) = \frac{f_{i,j}^{pgt}\left(g_k\right)^T}{\|g_k\| \cdot \|f_{i,j}^{pgt}\|} \tag{8}$$

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 $\max_{k \in N_c} \text{sim}\left(f_{i,j}^{pgt}, g_k\right)$. 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).

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 $x_i^u \in I_{diff}^u$:

$$S_{div}^i = 1 - \frac{1}{n_b^i} \sum_{j=1}^{n_b^i} \max_{k \in N_c} \text{sim}\left(f_{i,j}^{pgt}, g_k\right) \tag{9}$$

Higher S_{div}^i means more dissimilar the image is to other images, indicating that it may represent a novel class or a new aspect of a known class. Finally, we use *L-p* normalization to combine the two metrics as the final selection score of the unlabeled image x_i^u:

$$S_{sel}^i = \sqrt[p]{\left(S_{unc}^i\right)^p + \left(S_{div}^i\right)^p} \tag{10}$$

III. EXPERIMENT AND ANALYSIS

A. Datasets

To extensively evaluate the proposed framework, two representative and public datasets in remote-sensing images, known as DOTA [18] and DIOR [19] were employed in our experiments. DOTA [18] contains 2806 aerial images from different sensors and platforms with crowd sourcing and 188,282 instances, covered by 15 common object categories. DIOR [19] consists of 23,463 images and 192,472 instances that are manually labeled with axis-aligned bounding boxes, covering 20 object categories. In our experiments, we focus on the task of detection with horizontal bounding boxes (**HBB** for short). We divide both of the datasets into training set and test set, with a ratio of 2:1.

B. Design of the Experiments

In the experiments, similar to the previous works for SSOD [20], [16], the training set of datasets is randomly divided into the labeled set and the unlabeled set. The labeled set is initialized with 5% images from the training set of DOTA and 2.5% of DIOR. For evaluation, we adopt mAP (50:95) [21] as the basic comparison metric in our experiments. In the SSOD-AT method, we adopt Faster-RCNN with ResNet-50 as our basic detection network. 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 5% samples from the unlabeled data in DOTA and the 2.5% samples in DIOR for human labeling in each iteration.

TABLE I
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 II
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
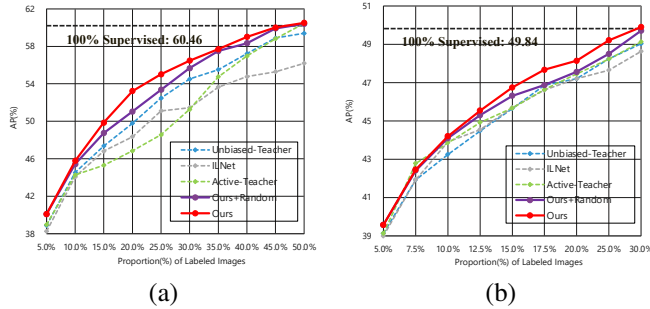
Fig. 2. Detection results of the different algorithms on the two remote-sensing datasets. (a) DOTA. (b) DIOR

TABLE III
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 IV
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**

C. Experimental Results and Analysis

1) Comparative Experiments With State-of-the-Art Methods: The proposed SSOD-AT is first evaluated on DOTA in Table I. Compared with the results of SSOD-AT(Random), our method consistently outperforms the comparison methods at each labeling ratio, which indicates RoICM is capable for detecting noisy RoIs in remote sensing images. Meanwhile, with the increasing of labeled images by active learning, the performance of our proposed method almost outperforms the SOTA methods in the whole active learning loops. From the results, we can observe that when Unbiased-Teacher and ILNet are adopted with the proposed active learning strategy, e.g. Unbiased-Teacher* and ILNet*, the performance in the active learning loops alway can achieve better performance, this further demonstrates the efficiency of our active learning strategy. The results are similar when we evaluate SSOD-AT on DIOR, as shown in Table II. In Fig. 2(a) and Fig. 2(b), our method achieves comparable detection capability to the fully supervised model with the label proportion of 50.0% on DOTA and 30% on DIOR, thereby substantially reducing the number of labeled samples for training.

2) Ablation Study: We ablate the two main parts of SSOD-AT, *RoICM* and *Sampling Strategy*, on DOTA and DIOR datasets in Table III and IV. In both tables, the first row represents a baseline teacher-student SSOD method(Unbiased-Teacher). The second row adds our novel RoICM module into the baseline with training samples randomly selected from datasets. The third and forth rows ablate the uncertainty and diversity modules in our sample strategy, respectively. It can be seen that the accuracy of SSOD-AT is gradually improved
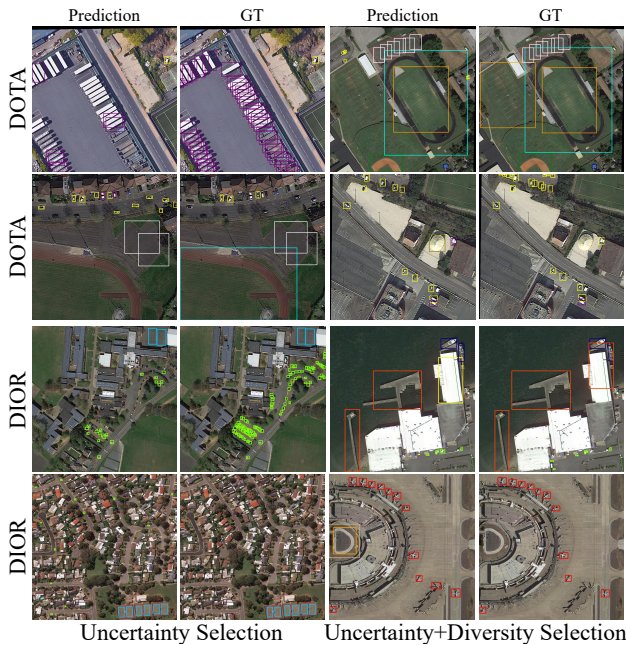
Fig. 3. Visualization of the images with top rank selection priority with different active sampling strategies. The Prediction columns denotes the pseudo-labels predicted by teacher network with 30%(DOTA) and 20%(DIOR) labeled proportions, while the GT columns refers to the corresponding ground-truths.

from adding RoICM to reject the noisy RoI to applying our sampling strategy to augment training samples.

3) Visualization Analysis: In Fig. 3, we visualize the examples selected by active sampling strategies when 30% images are labeled in DOTA and 20% images are labeled in DIOR. 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 missing since the detector is highly uncertain about these images. However, the samples selected by uncertainty are greatly susceptible to be category imbalanced (e.g., a significant fraction of images containing vehicles), while the addition of the diversity selection allows images containing more categories.

IV. CONCLUSION

We propose a novel method for semi-supervised object detection with active learning for remote sensing images. The proposed method 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 informative images from a vast pool of remote sensing images for expert labeling. Our proposed method is extensively evaluated on two remote sensing datasets, and it consistently outperforms state-of-the-art methods.

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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.