CLIP-GUIDED SOURCE-FREE OBJECT DETECTION IN AERIAL IMAGES

Nanqing Liu[1,2], *Xun Xu*[2], *Yongyi Su*[2], *Chengxin Liu*[2], *Peiliang Gong*[2], *Heng-Chao Li*[1]

[1]School of Information Science and Technology, Southwest Jiaotong University, PR China
[2]Institute for Infocomm Research (I2R), A-STAR, Singapore

ABSTRACT

Domain adaptation is crucial in aerial imagery, as the visual representation of these images can significantly vary based on factors such as geographic location, time, and weather conditions. Additionally, high-resolution aerial images often require substantial storage space and may not be readily accessible to the public. To address these challenges, we propose a novel Source-Free Object Detection (SFOD) method. Specifically, our approach is built upon a self-training framework; however, self-training can lead to inaccurate learning in the absence of labeled training data. To address this issue, we further integrate Contrastive Language–Image Pre-training (CLIP) to guide the generation of pseudo-labels, termed CLIP-guided Aggregation. By leveraging CLIP's zero-shot classification capability, we use it to aggregate scores with the original predicted bounding boxes, enabling us to obtain refined scores for the pseudo-labels. To validate the effectiveness of our method, we constructed two new datasets from different domains based on the DIOR dataset, named DIOR-C and DIOR-Cloudy. Experiments demonstrate that our method outperforms other comparative algorithms.

Index Terms— Source-free object detection, Aerial images, Self-training, CLIP, Domain adaptation

1. INTRODUCTION

In recent years, the field of object detection in aerial imagery [1, 2, 3, 4] has seen growing interest due to its relevance in areas such as urban planning, environmental monitoring, and disaster management. Deep learning methods have been particularly successful in aerial object detection. However, these methods usually require detailed instance-level annotations, which are both time-consuming and costly to obtain. Furthermore, these models often exhibit limited generalization when applied to aerial images taken under various conditions, such as using different sensors or in different weather, leading to issues related to domain gaps or dataset biases.

To address these challenges, unsupervised domain adaptive (UDA) object detection has become a promising solution. Yet, UDA still depends on labeled data from the source domain, which poses a challenge in aerial imagery. High-resolution aerial images typically require substantial storage space and may not be easily accessible to the public. To overcome these obstacles, source-free object detection (SFOD) [5] has been introduced. This approach relies solely on a pre-trained source model and an unlabeled target dataset, thus saving storage space and ensuring the security of sensitive data. Most current SFOD research [5] is based on self-training methods[6, 7], like the mean-teacher framework [8]. In this setup, a teacher model guides a student model, but there is a risk of error accumulation, known as *confirmation bias* if the teacher model provides incorrect learning targets.

Our solution aims to prevent the generation of inaccurate pseudo-label scores. Leveraging CLIP's [9] excellent domain adaptation features, we utilize CLIP to assist in generating pseudo-label scores across different domains. CLIP, which is known for learning transferable visual features from natural language supervision and has shown remarkable zero-shot classification abilities, is also effective in detection tasks. For instance, VFA [10] uses it as an aid for category judgments in few-shot object detection. Our approach, termed CLIP-guided Aggregation, combines the output scores of CLIP with those of the original teacher model. We compare labels generated by CLIP with the pseudo-labels from the teacher model. If they match, we keep the original classification scores; if they differ, we adjust the pseudo-label scores. This approach, using the CLIP model as an anchor [11] in the learning process, helps correct errors and reduce *confirmation bias*. Furthermore, the recent DOTA-C and DOTA-Cloudy dataset [12] introduced a variety of corruptions to assess the robustness of detectors. However, it requires testing on a server, a relatively cumbersome process. Based on this, we developed new datasets for different domains, DIOR-C and DIOR-Cloudy, from the DIOR dataset [3]. We also validated our method on DIOR-C and DIOR-Cloudy, effectively enhancing the performance of self-training-based SFOD, and achieving results significantly better than other comparative algorithms.

2. METHODOLOGY

In the context of Source-Free Object Detection (SFOD), practitioners are limited to using only the unlabeled target dataset, denoted as $\mathcal{D}_t$, and the pre-trained source model, Φ_p. Access to the labeled source dataset $\mathcal{D}_s$ is not available. As illustrated in Fig.1, our approach adopts a self-training method-

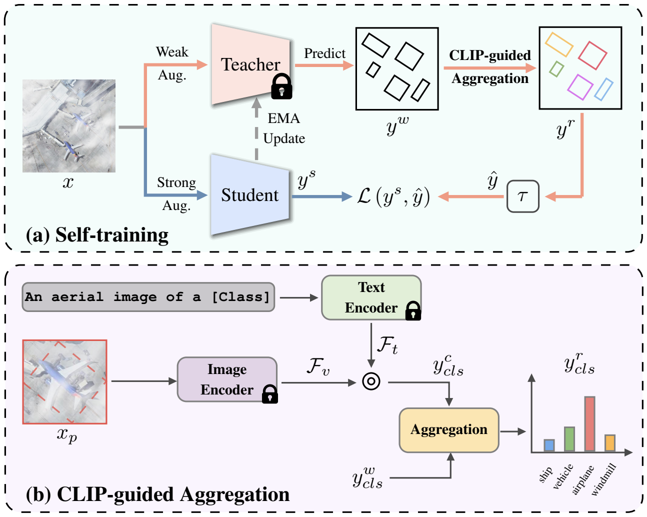


Fig. 1. Workflow of the proposed SFOD method. (a) Self-training framework. (b) CLIP-guided Aggregation.

ology [8] that is specifically adapted for the target dataset. During the pseudo-label generation phase, our method stabilizes the pseudo-label scores by incorporating a CLIP-guided Aggregation process. This section will primarily delve into these two key processes.

2.1. Self-training

In the SFOD task, for a randomly selected target image x from the dataset $\mathcal{D}_t$, we employ both weak and strong augmentation methods to generate x^w and x^s, respectively. Weak augmentation is limited to horizontal flipping with a probability of 0.5. On the other hand, strong augmentation encompasses a mix of color jittering, grayscale conversion, Gaussian blur, and the application of cutout patches. Both student and teacher models in our framework adopt the same network structure, specifically Oriented R-CNN [4].

As illustrated in Fig.1(a), for the weakly augmented images x^w, we input them into the teacher model Φ_t. Post-processing techniques such as Non-Maximum Suppression (NMS) are then applied to derive the classification scores y^w_{cls} and regression parameters y^w_{reg} for oriented bounding boxes. Recognizing the potential imprecision in initial object box scores, we implement a CLIP-guided Aggregation operation to refine the class scores, resulting in adjusted scores y^r_{cls}. The detailed methodology of this operation will be discussed in the following section. A confidence threshold τ is then used to filter out less probable predicted boxes, thus generating the pseudo-labels $\hat{y} = \{\hat{y}_{cls}, \hat{y}_{reg}\}$.

In the parallel process for the strongly augmented images x^s, these are fed into the student model Φ_s. This step similarly produces classification scores y^s_{cls} and regression parameters y^s_{reg} for oriented bounding boxes. The loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{RoI} + \mathcal{L}_{RPN} \tag{1}$$

where $\mathcal{L}_{RPN}$ and $\mathcal{L}_{RoI}$ represent the losses for the Region Proposal Network (RPN) and the Region of Interest (RoI) head, respectively.

To reduce the negative impact of inaccurate pseudo-labels, we update the teacher model through an exponential moving average (EMA) scheme, defined as $\Theta(\Phi_t) \leftarrow \alpha\Theta(\Phi_t) + (1-\alpha)\Theta(\Phi_s)$, with the coefficient α set at 0.998.

2.2. CLIP-guided Aggregation

When the teacher model generates predicted oriented boxes, it is crucial to refine their initial scores. For this, we utilize the zero-shot capabilities of CLIP [9] to assess the predicted patches. Since CLIP is designed to process horizontally oriented box images and our teacher model outputs aerial targets in rotated boxes, we first transform these rotated boxes into horizontal ones. Assuming the parameters of a rotated box are $\{x, y, w, h, \theta\}$, we calculate the width (w') and height (h') of the corresponding horizontal bounding box using the following equations: $w' = w \cdot |\cos\theta| + h \cdot |\sin\theta|, h' = w \cdot |\sin\theta| + h \cdot |\cos\theta|$.

As shown in Fig.1(b), we then use this horizontal bounding box to extract the relevant patch from the original image, denoted as $x_p \in \mathbb{R}^{N\times C\times H\times W}$. Here, N represents the number of patches, while C, H, and W denote the channels, height, and width of each patch, respectively. These patches x_p are fed into CLIP's image encoder to obtain the feature embeddings $\mathcal{F}_v \in \mathbb{R}^{N\times D}$, where D is the dimensionality of the embedding, typically 1024. Simultaneously, for the text branch and considering the aerial context, we formulate the text prompt as "`An aerial image of a [Class]`", with `[Class]` being a placeholder for the various detectable classes in the dataset. These prompts are processed by CLIP's pre-trained text encoder, resulting in embeddings $\mathcal{F}_t \in \mathbb{R}^{K\times D}$, where K is the number of classes. The category score $y^c_{cls} \in \mathbb{R}^{N\times K}$ for each patch, as determined by CLIP, is then calculated as $y^c_{cls} = \mathrm{Softmax}(\mathcal{F}_v \cdot \mathcal{F}_t^\top)$

In the process of refining the initial class scores predicted by the teacher model for each patch, denoted as $y^w_{cls} \in \mathbb{R}^{N\times K}$, we employ an aggregation method to derive refined scores $y^r_{cls} \in \mathbb{R}^{N\times K}$. The aggregation procedure is delineated as follows:

$$y^r_{cls} = \begin{cases} y^w_{cls}, & \text{if } \mathrm{argmax}(y^w_{cls}) = \mathrm{argmax}(y^c_{cls}), \\ (1-\lambda)y^w_{cls} + \lambda y^c_{cls}, & \text{otherwise.} \end{cases} \tag{2}$$

Note that argmax here represents the index corresponding to the maximum value. In this equation, the parameter λ serves to balance the original classification scores and those generated by CLIP. As indicated in Eq.2, when the category predicted by CLIP matches that of the original score, suggesting higher confidence in the score's accuracy, we retain the original score. Conversely, in instances of a mismatch, we adopt a weighted combination of both scores as the final score. This method effectively filters out unstable category scores and is independent of the teacher-student learning cycle, thereby reducing the likelihood of propagating incorrect labels.

Table 1. Source-free domain adaptive object detection results on **DIOR-C** and **DIOR-Cloudy** dataset.

Model	DIOR-C																			DIOR-Cloudy	mAP
	Ga.	Shot	Im.	Spec.	De.	Glass	Mo.	Zoom	Ga.	Snow	Frost	Fog	Br.	Spat.	Co.	El.	Pixel	JPEG	Sa.	Cloudy	
Clean	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54.1
Direct test	13.1	12.2	13.6	15.6	28.5	24.4	29.4	18.6	29.6	22.8	25.8	36.3	**49.9**	34.0	32.7	43.7	40.1	47.5	**52.6**	39.0	30.5
Tent[13]	15.8	14.6	16.4	16.6	19.5	16.4	21.9	11.9	20.6	16.3	18.2	26.8	28.5	21.6	27.5	26.1	26.5	25.8	30.4	23.6	21.3
BN[14]	18.1	17.3	18.7	19.7	22.6	18.4	24.2	14.3	23.0	19.1	20.4	30.4	32.0	31.3	30.8	29.1	30.2	28.9	30.4	26.7	24.3
Shot[15]	18.5	17.3	19.8	20.0	21.0	16.8	22.6	13.6	21.0	18.8	20.5	28.4	31.5	24.1	29.6	27.3	27.7	28.3	33.1	25.3	23.3
Self-training[8]	24.1	**21.1**	24.2	25.0	36.4	35.4	39.7	27.8	39.1	30.0	31.9	47.7	47.9	38.8	47.5	45.6	45.9	46.7	50.1	42.1	37.4
Ours	**25.6**	20.1	**26.0**	**25.4**	**37.9**	**35.9**	**40.2**	**30.3**	**40.1**	**31.5**	**32.7**	**50.0**	49.3	**39.8**	**47.8**	**47.3**	**47.0**	**47.6**	51.9	**44.1**	38.5

3. EXPERIMENTS

3.1. Experimental Setup

The datasets currently available for evaluating different aerial image domains, specifically DOTA-C[12] and DOTA-Cloudy[12], present certain limitations. Evaluations on these datasets require the use of DOTA's server and assessments for different types of corruption have to be conducted individually, posing significant challenges. To this end, we develop new datasets tailored for the SFOD task: DIOR-C and DIOR-Cloudy, derived from the publicly accessible DIOR dataset [3]. DIOR-C incorporates 19 types of corruptions from ImageNet-C [16], similar to those found in DOTA-C. For our experiments, we only generate images with a severity level of 3. DIOR-Cloudy is synthesized using publicly available cloudy images from the DOTA-Cloudy dataset. We use the original DIOR training set as our source data, and to create the unlabeled target training and test sets, we introduce corruptions to the original DIOR validation and test sets.

Our base network is the Oriented-RCNN with ResNet-50. In the source model training phase, we train on the source training set, initializing the feature backbone network with weights from a pre-trained ImageNet model. We set the batch size at 16 and used the SGD optimizer with a momentum rate of 0.9 and an initial learning rate of 0.01. The learning rate is decreased by a factor of 10 at the 24th and 33rd epochs, with the total training duration spanning 36 epochs. Following source domain training completion, the trained source model is used to initialize the teacher and student models for the self-training phase on the target training set. Here, the batch size is set to 2, and a single epoch of training is conducted using the SGD optimizer, with a momentum rate of 0.9 and a learning rate of 0.001. Experiments are performed using the PyTorch framework on a PC equipped with an Intel i7 single-core CPU and a GeForce RTX 3090 GPU. The mean Average Precision (mAP) on the target test set is reported, adhering to the PASCAL VOC evaluation metric.

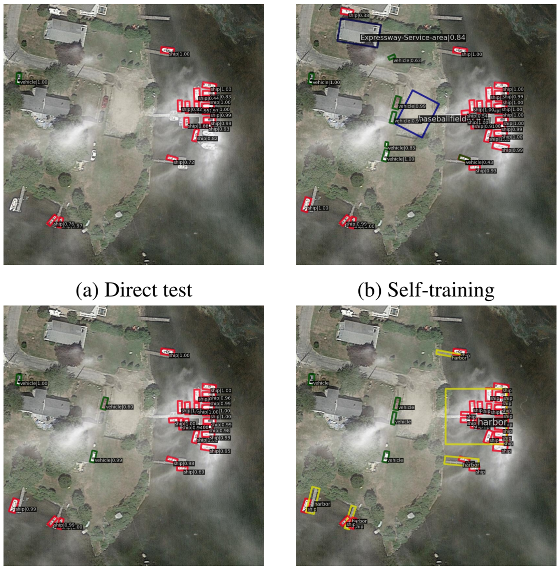
3.2. Competing Methods

We adopt the following generic state-of-the-art source-free domain adaptation methods to object detection tasks. **Direct test**: Directly apply the source model to test on the target dataset. **BN** [14]: Updates batch normalization statistics on the target data during testing. **Tent** [13]: Adapt a model by entropy minimization during testing. **Shot** [15]: Freeze the linear classification head and train the target-specific feature extraction module. **Self training** [8]: Two kinds of augmentation methods applying to target data. The pseudo-label is generated by the teacher network and used to supervise the student network.

We assess the performance of these methods on **DIOR-C** and **DIOR-Cloudy**, as detailed in Tab.1. The direct testing results show a significant decrease compared to the clean test (from 54.1% to 30.5%), highlighting the substantial impact of image corruption. Notably, several source-free domain adaptive classification methods, such as Tent, BN, and Shot, demonstrate limited effectiveness when directly applied to detection tasks. This shortcoming stems from classification methods relying on batch size-based statistics for adaptation, which is not feasible in detection scenarios due to typically smaller batch sizes. Consequently, these methods sometimes perform even worse than direct testing. However, self-training significantly improves outcomes (from 30.5% to 37.4%), suggesting its viability for SFOD tasks. Our proposed method, leveraging CLIP for more precise and stable pseudo-label generation, achieves the best results. We also visualize the results of different methods in Fig. 2. It can be observed that the direct test method exhibits instances of missed detections such as "*Vehicle*" and "*Ship*" because of domain shift. Although self-training can detect the required samples, it also introduces many false alarms, such as "*Baseball field*" and "*Expressway-Service-area*", due to the accumulation of errors during the training process. All methods overlooked the "*Harbor*" category in detection results, but ours closely approximates the ground truth.

3.3. Ablation Study

Our first experiment assesses the impact of varying λ values in Eq. 2. As depicted in Fig.3(left), we observe that the optimal performance is achieved when λ is set to 0.2. Increasing λ further leads to a decline in effectiveness, indicating that excessive reliance on CLIP can negatively affect the network's ability to generate accurate pseudo-labels. Additionally, we examine the influence of different CLIP encoder structures, as shown in Fig.3(right). The results suggest that CNNs generally outperform transformers in this context, mainly because the targets in aerial images are relatively sim-



(a) Direct test (b) Self-training

(c) Ours (d) Ground Truth

Fig. 2. Qualitative results of different methods on DIOR-Cloudy dataset.

ple. Consequently, the global information processing capabilities of transformers do not provide substantial benefits in these scenarios.

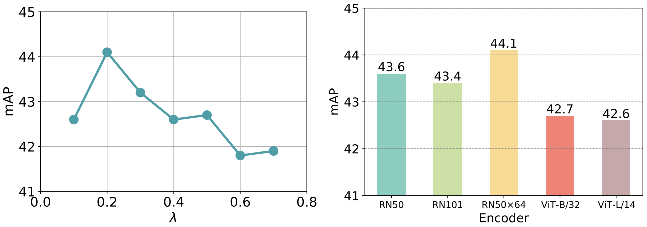


Fig. 3. Ablation experiments about different λ (left) and CLIP encoders (right) on the DIOR-Cloudy dataset.

4. CONLUSION

The paper presents a novel approach to source-free object detection in aerial images by integrating CLIP to guide pseudo-label score generation. The experiments were conducted on specially created datasets, DIOR-C and DIOR-Cloudy, derived from the publicly available DIOR dataset. The proposed method achieved the best results, outperforming other state-of-the-art source-free domain adaptation methods in object detection tasks.

5. REFERENCES

[1] Nanqing Liu, Turgay Celik, Tingyu Zhao, Chao Zhang, and Heng-Chao Li, “Afdet: Toward more accurate and faster object detection in remote sensing images,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 12557–12568, 2021.

[2] Nanqing Liu, Xun Xu, Turgay Celik, Zongxin Gan, and Heng-Chao Li, “Transformation-invariant network for few-shot object detection in remote-sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[3] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, 2020.

[4] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han, “Oriented r-cnn for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3520–3529.

[5] Vibashan VS, Poojan Oza, and Vishal M Patel, “Instance relation graph guided source-free domain adaptive object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3520–3530.

[6] Yongyi Su, Xun Xu, Tianrui Li, and Kui Jia, “Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering regularized self-training,” *arXiv preprint arXiv:2303.10856*, 2023.

[7] Yongyi Su, Xun Xu, and Kui Jia, “Towards real-world test-time adaptation: Tri-net self-training with balanced normalization,” *Proc. AAAI Conf. Artif. Intell.*, 2024.

[8] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda, “Unbiased teacher for semi-supervised object detection,” *Proc. Int. Conf. Learn. Represent.*, 2021.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8748–8763.

[10] Jiaming Han, Yuqiang Ren, Jian Ding, Ke Yan, and Gui-Song Xia, “Few-shot object detection via variational feature aggregation,” in *Proc. AAAI Conf. Artif. Intell.*, 2023.

[11] Yongyi Su, Xun Xu, and Kui Jia, “Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering,” *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 17543–17555, 2022.

[12] Haodong He, Jian Ding, and Gui-Song Xia, “On the robustness of object detection models in aerial images,” 2023.

[13] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell, “Tent: Fully test-time adaptation by entropy minimization,” *Proc. Int. Conf. Learn. Represent.*, 2021.

[14] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. Int. Conf. Mach. Learn.* pmlr, 2015, pp. 448–456.

[15] Jian Liang, Dapeng Hu, and Jiashi Feng, “Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation,” in *Proc. Int. Conf. Mach. Learn.* PMLR, 2020, pp. 6028–6039.

[16] Dan Hendrycks and Thomas Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” *Proc. Int. Conf. Learn. Represent.*, 2019.