

# LLM-Assisted Semantic Guidance for Sparsely Annotated Remote Sensing Object Detection

Wei Liao<sup>1</sup>, Chunyan Xu<sup>1\*</sup>, Chenxu Wang<sup>1</sup>, Zhen Cui<sup>2</sup>

<sup>1</sup>Nanjing University of Science and Technology, Nanjing, Jiangsu, China

<sup>2</sup>Beijing Normal University, Beijing, China

## Abstract

Sparse annotation in remote sensing object detection poses significant challenges due to dense object distributions and category imbalances. Although existing Dense Pseudo-Label methods have demonstrated substantial potential in pseudo-labeling tasks, they remain constrained by selection ambiguities and inconsistencies in confidence estimation. In this paper, we introduce an LLM-assisted semantic guidance framework tailored for sparsely annotated remote sensing object detection, exploiting the advanced semantic reasoning capabilities of large language models (LLMs) to distill high-confidence pseudo-labels. By integrating LLM-generated semantic priors, we propose a Class-Aware Dense Pseudo-Label Assignment mechanism that adaptively assigns pseudo-labels for both unlabeled and sparsely labeled data, ensuring robust supervision across varying data distributions. Additionally, we develop an Adaptive Hard-Negative Reweighting Module to stabilize the supervised learning branch by mitigating the influence of confounding background information. Extensive experiments on DOTA and HRSC2016 demonstrate that the proposed method outperforms existing single-stage detector-based frameworks, significantly improving detection performance under sparse annotations. Our source code is available at <https://github.com/wuxiuzhilianni/RSST>.

## 1. Introduction

Object detection has been a pivotal task in computer vision, underpinning a wide spectrum of real-world applications, including autonomous driving, surveillance, and remote sensing analysis. Traditional object detection models require extensive human-annotated datasets to ensure optimal performance [6, 18]. To mitigate this reliance on fully annotated data, *semi-supervised object detection* (SSOD) [2, 17, 32] has emerged as a promising paradigm, leveraging

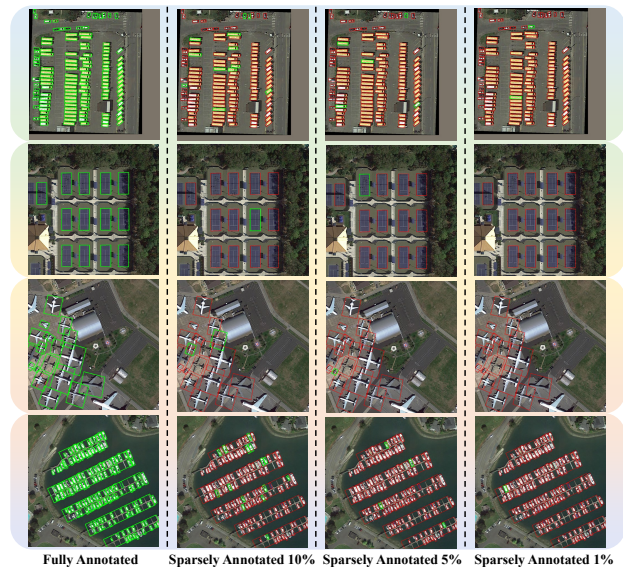


Figure 1. The illustration of sparsely annotated objects in aerial images. From left to right, the columns represent the fully annotated setting, as well as sparsely annotated settings with 10%, 5%, and 1% of the original annotations, respectively. Green bounding boxes indicate the ground truth annotations in the original dataset, while red bounding boxes denote annotations that have been intentionally removed during training.

a limited set of fully labeled images alongside an abundant corpus of unlabeled images to enhance model generalization while alleviating annotation burdens. Despite its success in generic vision tasks, SSOD remains underexplored in the domain of remote sensing imagery [11, 14, 37]. This discrepancy stems from the intrinsic challenges posed by remote sensing data, wherein objects are often densely distributed, exhibit significant intra-class variations, and manifest diverse orientations across images. To address this limitation, researchers have introduced the concept of *sparsely annotated object detection* (SAOD) [25, 28, 31], where only a fraction of objects within an image are explicitly labeled, while the rest remain unannotated. Fig. 1 presents a schematic illustration of sparse annotations with varying label rates for remote sensing images.

\*Corresponding Author: C. Xu (cyx@njust.edu.cn).

Pseudo-labeling [13], in conjunction with consistency-based regularization [1], forms a cornerstone of SSOD. In this paradigm, the teacher detector generates pseudo-labels, which serve as supervisory signals for training the student model on unlabeled data, thereby reducing reliance on manually annotated datasets. Some methods adopt pseudo-boxes, where the predicted bounding boxes undergo multiple refinement procedures and serve as direct supervision for the student detector [22, 23, 41]. However, these additional post-processing steps introduce potential information loss and may lead to suboptimal supervision quality. Alternatively, other approaches employ Dense Pseudo-Labels (DPL) by leveraging raw model outputs without conventional post-processing [20, 34, 47], which can provide richer supervision signals and potentially enhance detection performance under limited annotations [15].

Prior methodologies leveraging DPL predominantly employ a top- $k$  selection mechanism to identify foreground pixels for distillation. However, this heuristic approach introduces inherent selection and assignment ambiguities, which undermine the effectiveness of pseudo-labeling. To alleviate these ambiguities, recent works have explored alternative selection paradigms. For instance, ARSL [20] refines pseudo-label selection by jointly considering classification confidence and localization quality, whereas MCL [34] adopts a divide-and-rule strategy, wherein distinct selection rules are formulated for objects of varying scales. Despite these improvements, a fundamental inconsistency arises between the mean confidence of pseudo-labels and their associated categorical assignments. Specifically, some underrepresented categories exhibit relatively elevated confidence scores, leading to their frequent selection during the assignment process. In contrast, categories with larger sample sizes tend to display markedly lower confidence, resulting in inadequate model training. This misalignment between localized confidence estimates and the true underlying semantic structures introduces substantial ambiguity, thereby disrupting the model’s learning dynamics.

To address the aforementioned challenges, we propose an LLM-assisted semantic guidance framework grounded in a Multi-Branch Input (MBI) architecture, which synergistically integrates supervised and unsupervised learning paradigms to enhance semantic consistency and pseudo-label reliability. Within the supervised learning branch, the Adaptive Hard-Negative Reweighting (AHR) module is designed to dynamically modulate the influence of hard negatives in the logit space, thereby mitigating their adverse effects on model optimization. Concurrently, in the unsupervised learning branch, the LLM-Assisted Semantic Prediction (LSP) module autonomously generates and gradually refines class-specific prompts, ensuring the progressive enhancement of semantic representations. These prompts subsequently support the Class-Aware Label As-

ignment (CLA) mechanism, which adaptively modulates the assignment strategies to accommodate both fully unlabeled and sparsely labeled data distributions. By incorporating these synergistic components, the proposed framework effectively mitigates the misalignment between localized confidence estimates and the underlying semantic structures, ultimately fostering a more robust and semantically coherent learning paradigm.

To summarize, the principal contributions of this work are as follows: i) We propose an LLM-assisted semantic guidance framework, meticulously tailored to accommodate the challenges posed by the sparsely annotated object detection (SAOD) task. ii) To alleviate the misalignment between localized confidence estimates and the underlying semantic structures, we have constructed several effective modules for both the supervised and unsupervised branches. iii) We conduct comprehensive experiments on two remote sensing datasets, including DOTA [39] and HRSC2016 [24], demonstrating the effectiveness and robustness of our model.

## 2. Related Work

### 2.1. Oriented Object Detection

Oriented object detection has emerged as a vital research area in computer vision, addressing the challenges posed by objects with arbitrary orientations, particularly in aerial image analysis. Unlike traditional object detectors that rely on horizontal bounding boxes (HBB) [7, 29, 30, 33, 46], oriented object detection techniques utilize oriented bounding boxes (OBB), which provide a more accurate representation of rotated objects [8, 40, 43]. In recent years, several innovative approaches have been developed to improve the accuracy and robustness of oriented object detection. CSL [42] reformulates the angle regression problem as a classification task to address the issue of discontinuous boundaries caused by angular periodicity or corner ordering, introducing a circular smooth label technique to improve error tolerance and enhance detection performance. S2A-Net [9] leverages the feature alignment module to generate high-quality anchors, followed by the oriented detection module, which employs active rotating filters to produce features that are both orientation-sensitive and orientation-invariant. In contrast to the above studies, which primarily focus on the supervised learning paradigm, this work takes an initial step towards exploring sparsely annotated oriented object detection, aiming to reduce annotation costs and improve detector performance by utilizing unlabeled data.

### 2.2. Label-Efficient Object Detection

Label-efficient object detection aims to mitigate the dependency on fully annotated datasets by leveraging weak supervision, self-training, and pseudo-labeling techniques. Semi-supervised object detection (SSOD) has advanced

with teacher-student frameworks, where the teacher model, updated via exponential moving average (EMA), generates pseudo-labels to guide student training [21, 38, 44]. Soft Teacher [41] employs box jittering for pseudo-box refinement, while Unbiased Teacher [22] and its improved variant [23] address class imbalance and enhance anchor-free detectors, respectively. Dense Teacher [47] introduces dense pseudo-labels to replace heuristic post-processing steps. ARSL [20] integrates joint-confidence estimation and task-separation assignment to improve label assignment at the pixel level. Sparsely annotated object detection (SAOD) addresses scenarios where only a fraction of instances are labeled, reducing annotation costs while maintaining detection performance. The absence of complete annotations introduces challenges, such as missing supervision signals. Niitani et al. [26] propose part-aware sampling to mitigate incorrect supervision. Co-mining [36] employs a Siamese network to enhance multi-view learning and better mine unlabeled instances. Region-based approaches [28] treat SAOD as a semi-supervised problem, identifying unlabeled foreground regions. Calibrated Teacher [35] refines confidence estimation to stabilize training, ensuring consistent pseudo-label quality. PECL [25] introduces a progressive selection strategy tailored to aerial images. In contrast to prior methods, our approach innovatively incorporates dense label representations into sparsely annotated aerial object detection, refining optimization under minimal supervision by integrating self-training and leveraging inter-class relational priors to enhance detection robustness.

### 3. Method

#### 3.1. Overview

SAOD presents significant challenges due to limited supervisory signals, necessitating the effective utilization of sparsely annotated instances and unlabeled images to train a robust detector capable of accurately identifying all valid targets. Formally, the sparsely annotated training set is defined as  $\mathcal{X}_t = \{(x_i, y_i)\}_{i=1}^N$ , where  $N$  represents the total number of images in the training dataset. For each image, denoted as the  $i$ -th sample, its sparse annotations are represented as  $y_i = \{(c_i^j, \theta_i^j, b_i^j)\}_{j=1}^{N_{il}}$ , where  $c_i^j$ ,  $\theta_i^j$ , and  $b_i^j$  correspond to the class label, orientation, and bounding box coordinates of the  $j$ -th annotated object, respectively. In this context,  $N_{il}$  denotes the number of labeled instances present in the  $i$ -th image, with  $N_{il} \geq 0$  indicating the possibility of zero labeled instances.

Fig. 2 presents an overview of the LLM-assisted semantic guidance framework. The framework is designed as a self-supervised, single-stage detection paradigm based on the Multi-Branch Input (MBI) Architecture, where a tightly coupled, cyclic optimization mechanism coordinates the reciprocal refinement between the generation of conformal

dense pseudo-labels and the progressive adaptation of the detector. The framework comprises a supervised and an unsupervised branch, each playing a distinct role in optimizing the model’s learning process. In the supervised branch, the model is trained with sparsely annotated data, providing the primary supervisory signal that guides the optimization trajectory of the student detector throughout the training process. During this phase, the Adaptive Hard-Negative Reweighting (AHR) module is employed to regulate hard negatives, facilitating a more discriminative learning process. In the unsupervised branch, the LLM-Assisted Semantic Prediction (LSP) module utilizes a large language model (LLM) to autonomously generate and progressively refine class-specific prompts for each image in an offline manner. This refinement significantly enhances the subsequent Class-Aware Label Assignment (CLA) mechanism, which strategically adapts the assignment strategies based on both labeled and sparsely labeled data. This approach mitigates foreground class imbalance and strengthens the overall robustness of the detection framework.

#### 3.2. LLM-Assisted Semantic Prediction

Prevailing strategies employing DPL for model training primarily delineate label assignment protocols by identifying the top- $k$  most confident predictions either across the global spatial extent [47], or within heterogeneous feature maps through a divide-and-rule paradigm [34]. However, such approaches intrinsically overlook a pivotal aspect: the explicit semantic determination of foreground categories. As illustrated in Fig. 3, certain underrepresented categories exhibit relatively high confidence scores, leading to their frequent selection during the assignment process. Conversely, some categories with a larger sample count tend to have significantly lower confidence scores, resulting in insufficient training for the model. The misalignment between localized confidence estimations and the true underlying semantic structures introduces substantial ambiguity into the learning dynamics.

Drawing inspiration from [12], we address this challenge by employing a large language model (LLM) to assist offline pseudo-label assignment before the training stage. For each image, the LLM is leveraged to infer potential foreground categories, which are subsequently utilized to enhance the unsupervised label assignment. Specifically, given the complete set of categories in the dataset, denoted as  $\mathcal{C} = \{class_1, class_2, \dots, class_{15}\}$  in DOTA, along with an additional background category  $\emptyset$ , we formulate a structured linguistic query to guide the LLM in foreground category inference. The designed instruction is as follows: “Choose categories presented in the image:  $class_1, \dots, class_{15}, none$ . Choose one or several classes. Answer in one word or a short phrase.” Next, high-

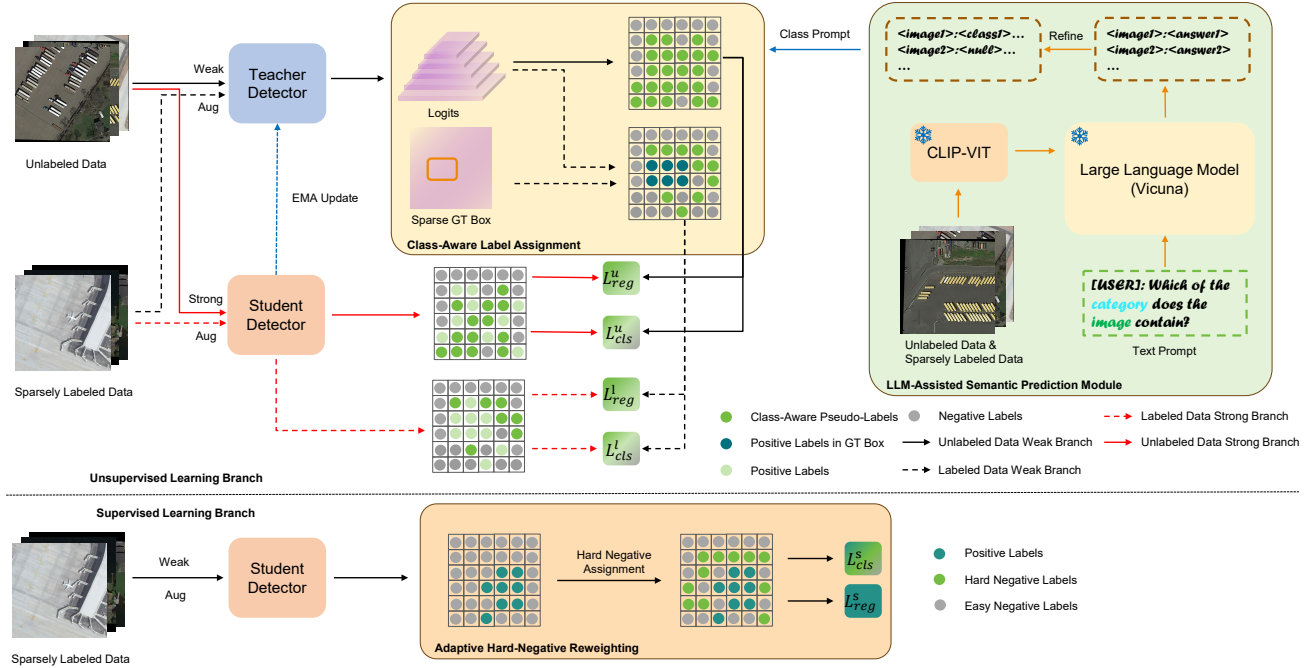


Figure 2. The illustration of the LLM-assisted semantic guidance framework. In the unsupervised branch, the LLM-Assisted Semantic Prediction (LSP) module generates and refines class prompts, while the Class-Aware Label Assignment (CLA) selects high-quality pixel-level pseudo-labels using distinct strategies for sparsely labeled and unlabeled data. In the supervised branch, the Adaptive Hard-Negative Reweighting (AHR) module identifies hard negatives in the logit space and mitigates their impact by reducing weights.

resolution remote sensing imagery is processed through CLIP-ViT (L-14) [27], where the positional encoding is interpolated to accommodate an expanded input resolution of  $504 \times 504$ , increasing the number of patches to 1296. An MLP-based adaptor then maps the 1024-dimensional visual embeddings into a 4096-dimensional space, ensuring compatibility with the Vicuna-v1.5 (7B) model [4]. The structured visual representations, concatenated with the formulated query, are fed into the LLM to facilitate semantically coherent category identification. Leveraging its contextual reasoning within an autoregressive decoding paradigm, the LLM derives the most probable foreground categories. If the model does not identify any foreground objects, it outputs `none`. Once foreground predictions are obtained for all images, a refinement step is applied: for sparsely annotated data, where labeled instances already contain foreground category information, the final category set is derived by taking the union of the LLM-generated predictions and the provided annotations. In contrast, for unlabeled data, the predictions remain unchanged. This refinement strategy ultimately ensures that approximately four-fifths of the final predictions are correct.

### 3.3. Class-Aware Label Assignment

With the assistance of the class prompt generated by LSP, foreground category information is explicitly leveraged to refine the assignment paradigm of positive and negative

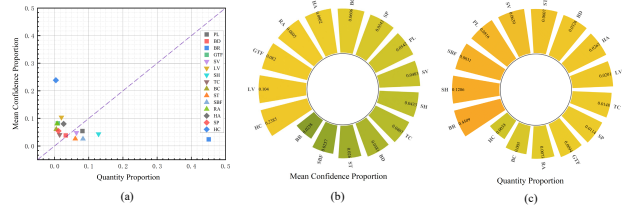


Figure 3. Analysis of Imbalance Between Category-Wise Quantity Proportion and Mean Confidence Proportion. (a) A scatter plot illustrating the relationship between Quantity Proportion (x-axis) and Mean Confidence Proportion (y-axis) for each category. (b) A radial bar chart displaying the Quantity Proportion of each category in ascending order, with category labels annotated accordingly. (c) A radial bar chart presenting the Mean Confidence Proportion of each category, sorted in ascending order, with category labels provided.

samples within the unsupervised learning distillation process. By embedding class-specific priors into the distillation framework, the proposed methodology mitigates the intrinsic ambiguities associated with high-confidence pseudo-label predictions, thereby enhancing their semantic reliability and robustness. To optimize the selection of supervisory signals, distinct assignment strategies are employed for different data types. To further handle fully unlabeled data, we propose a label assignment method that leverages the class prompt generated by the LSP module. This method aims to select a fixed number of foreground pixels through a struc-

tured three-stage process, ensuring that the selected pixels exhibit high semantic relevance and are well-distributed across different foreground categories.

Given a pixel set  $P = \{p_i\}$ , we prioritize selecting pixels that are semantically associated with foreground categories, as specified by a predefined class prompt. This ensures that the selected pixels are aligned with the intended semantic content. To further enhance the reliability of the selection process and suppress potentially noisy or ambiguous samples, pixels with confidence scores falling below a predefined threshold  $thr$  are systematically filtered out:

$$P_{fg} = \{p \in P \mid C(p) \in \text{Prompt} \wedge S(p) > thr\}, \quad (1)$$

where  $C(p)$  denotes the predicted category of pixel  $p$ , and  $S(p)$  represents the joint confidence score of pixel  $p$ , as defined in [20]. The threshold  $thr$  serves as a confidence filter, removing ambiguous pixels with the low reliability. To enhance generalization, we introduce an additional selection mechanism based on the pixel confidence. Specifically, we select the top- $k$  pixels with the highest joint confidence scores:

$$P_{conf} = \text{Top}_k\left(\sum_{p \in P} S(p)\right), \quad (2)$$

where  $\text{Top}_k(\cdot)$  denotes the operation of selecting the highest- $k$  scoring pixels. To ensure balanced representation across different foreground categories, we further refine the pixel selection by choosing the top- $k_j$  pixels with the highest confidence scores within each individual category  $c_j$ :

$$P_{c_j} = \text{Top}_{k_j}\left(\sum_{p \in P} S_j(p)\right), \quad (3)$$

where  $S_j(p)$  represents the confidence measure of the pixel  $p$  corresponding to the category  $c_j$ . Ultimately, to eliminate redundancy and enhance the heterogeneity of the selected pixels, a deduplication operation is conducted. The resulting final set of selected pixels from the unlabeled images is defined as follows:

$$P_{\text{unlabeled}} = \text{Unique}\left(P_{fg} \cup P_{conf} \cup \bigcup_j P_{c_j}\right), \quad (4)$$

where  $\text{Unique}(\cdot)$  ensures that each selected pixel appears only once in the final set. This final selection process guarantees a diverse and semantically meaningful assignment of pixels, facilitating more robust learning.

To address sparsely labeled data, we propose an assignment strategy that incorporates additional supervisory signals from available annotations, including both the class prompts generated by the LSP module and the ground truth information. Given the set of annotated pixels  $P_{GT}$ , we directly include these pixels as positive samples:

$$P_{GT} = \{p \in P \mid p \in \mathcal{A}\}, \quad (5)$$

where  $\mathcal{A}$  represents the set of manually labeled pixels. By incorporating these reliable ground-truth (GT) pixels, we ensure that the model receives explicit supervision from human-labeled data, thereby mitigating potential errors introduced by the pseudo-labeling process. The final selection of positive samples combines four sources: (1) pixels aligned with the class prompt, (2) high-confidence pixels, (3) category-balanced pixels, and (4) ground-truth pixels. To eliminate redundancy and maintain sample diversity, a deduplication process is applied:

$$P_{\text{sparsely}} = \text{Unique}\left(P_{fg} \cup P_{conf} \cup \bigcup_j P_{c_j} \cup P_{GT}\right), \quad (6)$$

where  $\text{Unique}(\cdot)$  removes duplicate entries, ensuring that each selected pixel appears only once in the final set. This strategy effectively enhances the supervision quality, enabling more reliable training under sparse annotations.

### 3.4. Adaptive Hard-Negative Reweighting

One-stage detectors are inherently more vulnerable to incomplete or missing annotations [16, 19, 45], particularly in the SAOD task, where distinguishing hard negative samples poses a significant challenge. These hard negatives may correspond to genuinely unannotated objects or background clutter, leading to increased uncertainty and hindering model convergence.

To address this challenge, we propose AHR to adaptively modulate the loss contribution of negative samples based on their confidence scores. This strategy aims to strike a balance between mitigating the impact of erroneous supervision and preserving beneficial background cues. The loss function is formulated as follows:

$$L(p_t) = \begin{cases} -\log(p_t)\alpha p_t^\gamma, \\ -\log(1-p_t)(1-\alpha)(1-p_t)^\gamma, \\ -\log(1-p_t)(1-\alpha)(1-p_t)^\gamma w, \end{cases} \quad (7)$$

where  $p_t$  represents the predicted confidence score,  $\alpha$  controls the weighting of positive and negative samples, and  $\gamma$  adjusts the focusing effect to suppress well-classified samples. The adaptive scaling factor  $w$  down-weights misleading high-confidence negatives. The three cases in the loss function correspond to: positive samples, negative samples with confidence below a threshold  $thr$ , and negative samples with confidence above  $thr$ .

## 4. Experiments

### 4.1. Experimental Setup

**Datasets:** We evaluate our approach on two widely used aerial object detection benchmarks: DOTA [39] and HRSC2016 [24]. The DOTA dataset consists of 2,806 high-resolution aerial images with 188,282 annotated instances

exhibiting significant variations in scale, aspect ratio, and orientation. The dataset is partitioned into three subsets: a training set containing 1411 images, a validation set with 458 images, and a test set comprising 937 images. It encompasses 15 distinct object categories, namely plane (PL), baseball-diamond (BD), bridge (BR), ground-track-field (GTF), small-vehicle (SV), large-vehicle (LV), ship (SH), soccer-ball field (SBF), tennis-court (TC), basketball-court (BC), storage-tank (ST), roundabout (RA), harbor (HA), swimming-pool (SP), and helicopter (HC). Under sparse annotation settings, we randomly sample 1%, 2%, 5%, and 10% of instances per category before tiling each image into non-overlapping  $1024 \times 1024$  patches. The HRSC2016 dataset contains 436 training, 181 validation, and 444 test images. Following [25], we construct sparsely labeled variants by retaining 10% of annotations, with at least one per category per image. The performance is evaluated using the standard mean average precision (mAP).

**Implementation Details:** Our proposed models are implemented and trained using the MMDetection [3] and MM-Rotate [48] frameworks. Within this framework, we employ the Rotated FCOS [33] as the single-stage detection paradigm, leveraging a ResNet-50 [10] backbone pretrained on ImageNet [5] to extract hierarchical feature representations. We employ Stochastic Gradient Descent (SGD) as the optimization algorithm, with an initial learning rate of 0.0025, a momentum coefficient of 0.9, and a weight decay factor of 0.0001 to prevent overfitting. The total training schedule comprises 12,000 iterations, with an initial burn-in phase of 6,400 iterations to stabilize model convergence.

Table 1. Performance comparison of the OBB task under different annotation rates (1%, 2%, 5%, and 10%) on the DOTA dataset. Only the main mAP results are reported. The detection methods are based on Rotated-FCOS (†), Rotated-RetinaNet (‡) and Faster-RCNN (○).

Model Type	Model Name	1%	2%	5%	10%
Supervised	Rotated FCOS	41.18	43.46	48.18	49.79
	Rotated RetinaNet	43.12	45.09	48.86	53.16
	S <sup>2</sup> A-Net	40.14	41.92	48.78	54.39
	Rotated Faster R-CNN	45.44	48.31	51.81	57.75
	Oriented R-CNN	51.05	53.09	56.85	61.50
	ReDet	50.72	52.08	58.01	61.79
Semi-Supervised	Unbiased Teacher <sup>○</sup>	43.59	44.97	51.55	57.23
	Dense Teacher <sup>†</sup>	-	47.72	50.82	58.13
	PseCo <sup>○</sup>	43.03	46.10	52.27	57.76
	SOOD <sup>‡</sup>	45.34	47.60	52.84	57.36
	ARSL <sup>†</sup>	43.88	44.39	51.38	55.61
	MCL <sup>†</sup>	41.90	43.75	51.53	56.17
Sparsely-Annotated	S <sup>2</sup> A-Net w/PECL <sup>†</sup>	50.39	53.81	57.42	62.49
	Ours <sup>†</sup>	56.64	59.37	63.95	65.11

## 4.2. Experimental Results

**DOTA:** We report the performance of our method on the oriented bounding box (OBB) detection task using the

Table 2. Prediction Statistics of LLM-Assisted Semantic Prediction Module. The total number of images is 21,046. “None” indicates that both the prediction and ground truth (GT) have no foreground classes. “Exact” means the predicted foreground classes are identical to the GT. “Partly” denotes cases where the predicted classes are a subset of the GT.

Modification (%)	Correct Predictions			Errors
	None	Exact	Partly	
0 (Original)	6943	4861	3771	5471
1	6819	7539	2259	4429
2	6821	7755	2139	4331
5	6829	8075	1976	4166
10	6817	8447	1765	4017

Table 3. Prediction Metrics and Cost for Class Prompts Synthesized by Distinct Large Language Models.

Model	None	Exact	Partly	Errors	Params(B)	T/img (s)
Phi-3.5-Vision-Instruct	2096	3036	3015	12899	4.15	0.85
Qwen-VL-Chat-Int4	2933	2292	2097	13724	3.18	0.82
InternVL2.5-1B	2697	3177	3479	11693	0.94	0.39
InternVL2.5-2B	1752	3906	3077	12311	2.21	0.44
InternVL2.5-4B	2169	4045	3506	11326	3.71	0.61
LLaVA-v1.5-7B	4485	4182	3446	8933	7.06	0.26
Ours	6943	4861	3771	5471	7.06	0.48

Table 4. Comparative analysis of prediction consistency under diverse textual prompt formulations within the LLM-Assisted Semantic Prediction module.

Prompt ID	None	Exact	Partly	Errors
P1	-	682	203	20161
P2	-	5711	4332	11003
P3	1224	3560	3093	13169
P4	6979	4616	3794	5657
P5	6943	4861	3771	5471

### Prompt Descriptions:

P1: Which of the following categories does the image contain: plane, ..., none.

P2: Which of the following categories does the image contain: plane, ..., helicopter. Answer in one word or a short phrase.

P3: Please classify this image among: plane, ... , none.

P4: What objects are in this image? Choose from the following: plane, ..., none. Answer in one word or a short phrase.

P5: Which of the following categories does the image contain: plane, ..., none. Answer in one word or a short phrase.

DOTA benchmark, as detailed in Table 1. Overall, in comparison to existing methodologies under fully supervised, semi-supervised, and sparse annotation regimes, the proposed framework demonstrates pronounced and consistent improvements in detection precision across all annotation ratios. In particular, relative to Dense Teacher [47], our method achieves significant gains of 11.65%, 13.13%, and 6.98% under 2%, 5%, and 10% label rates, respectively. Furthermore, we conduct a comprehensive statistical analysis of the accuracy of class prompts generated by LSP module for the DOTA dataset, as presented in Table 2. The empirical results indicate that an increasing label rate effectively contributes to a reduction in the number of incor-

Table 5. Quantitative comparison of representative oriented object detection methods on the HRSC2016 dataset under the Oriented Bounding Box (OBB) setting. The performance is evaluated in terms of mAP, as well as AP at 50% and 75% IoU thresholds.

Method	Performance (%) $\uparrow$		
	mAP	AP50	AP75
Rotated RetinaNet	47.96	82.70	49.90
S <sup>2</sup> A-Net	48.05	79.60	51.60
Rotated FCOS	49.85	79.70	56.10
Oriented RCNN	57.26	80.90	68.70
ReDet	63.80	88.50	77.50
Dense Teacher	64.86	88.00	75.60
Ours	<b>66.80</b>	88.40	77.40

rect prompts. This improvement arises from the capability of leveraging annotation information embedded within the labeled data to iteratively refine the class prompts generated by the large language model. Additionally, as delineated in Table 3, the LLM adopted in our method is specifically adapted for remote sensing, yielding superior inference performance with competitive latency compared to generic LLMs trained on conventional datasets. Furthermore, Table 4 presents a comparative study on various textual prompt designs. The prompt adopted in our work was selected through extensive empirical tuning to maximize task-specific performance.

**HRSC2016:** To further substantiate the efficacy of our proposed framework, we conduct comprehensive comparative analyses on the HRSC2016 dataset. The quantitative performance comparisons between our framework and various detector baselines are systematically presented in Table 5. Specifically, we report the mean Average Precision (mAP) metric, as well as the AP<sub>50</sub> and AP<sub>75</sub> scores, which serve as standard benchmarks for evaluating detection accuracy. Notably, when benchmarked against the dense teacher [47], our framework exhibits substantial improvement of 1.94% in terms of mAP. This empirical evidence underscores the robustness and adaptability of our approach, even when evaluated on relatively small-scale aerial image datasets.

### 4.3. Ablation Study

We conduct comprehensive ablation experiments to evaluate the performance of our framework under different settings. All the experiments are conducted on the DOTA dataset at the 5% label rate, unless stated otherwise.

As illustrated in Table 6, we conduct a rigorous ablation study to systematically evaluate the contribution of each proposed component. The experimental findings reveal that even when solely employing the AHR, our approach already surpasses the baseline, highlighting the effectiveness of adaptive negative sample in improving detection performance. Furthermore, the incorporation of the LSP, CLA and

Table 6. Ablation study comparing the mean Average Precision (mAP) performance resulting from the integration of various architectural components within the proposed framework. The evaluation is conducted under multiple supervision ratios (2%, 5%, and 10%) to assess the effectiveness and generalization capability of each configuration.

	Components				mAP (%) $\uparrow$		
	AHR	LSP	CLA	MBI	2%	5%	10%
<b>Baseline</b>	×	×	×	×	47.72	50.82	58.13
<b>Ours</b>	✓	×	×	×	55.38	61.60	62.93
	✓	✓	✓	×	58.73	63.65	64.77
	✓	✓	✓	✓	59.37	63.95	65.11

Table 7. Comparative Analysis of Assignment Mechanisms on the DOTA dataset under Uniform Experimental Protocols.

Assignment Strategy	2%	5%	10%
SOOD	54.38	60.02	63.27
Dense Teacher	55.38	61.60	62.93
MCL	57.63	62.11	63.27
Ours (no-prompt)	56.17	61.97	63.53
Ours (LLM-guided prompt)	58.73	63.65	64.77
Ours (gt-prompt)	60.45	65.04	66.41

Table 8. Quantitative evaluation of the proposed Adaptive Hard-Negative Reweighting (AHR) module under varying configurations of weighting coefficients and prediction confidence thresholds on the training dataset.

Weight	Threshold		
	1.0	0.95	0.90
0.30	67.87	67.53	67.76
0.25	66.82	67.81	67.70
0.20	67.89	68.01	67.90
0.15	67.31	67.69	<b>68.20</b>
0.10	67.83	67.73	68.01

MBI modules provides additional performance gains, substantiating their respective roles in enhancing feature representation learning and refining pseudo-label assignments. These results collectively underscore the efficacy of our proposed framework in leveraging complementary strategies to achieve superior detection accuracy under limited supervision. Furthermore, Fig. 4 provides a visualization of the foreground pixels selected by our model during training. Notably, the distribution of the identified positive samples aligns closely with the locations of ground truth bounding boxes in the original images, further validating the effectiveness of our assignment strategy in accurately capturing salient object regions.

To address the premise that erroneous predictions may impart a non-trivial perturbation to convergence dynamics, our assignment framework integrates three foundational mechanisms deliberately formulated to suppress such detri-

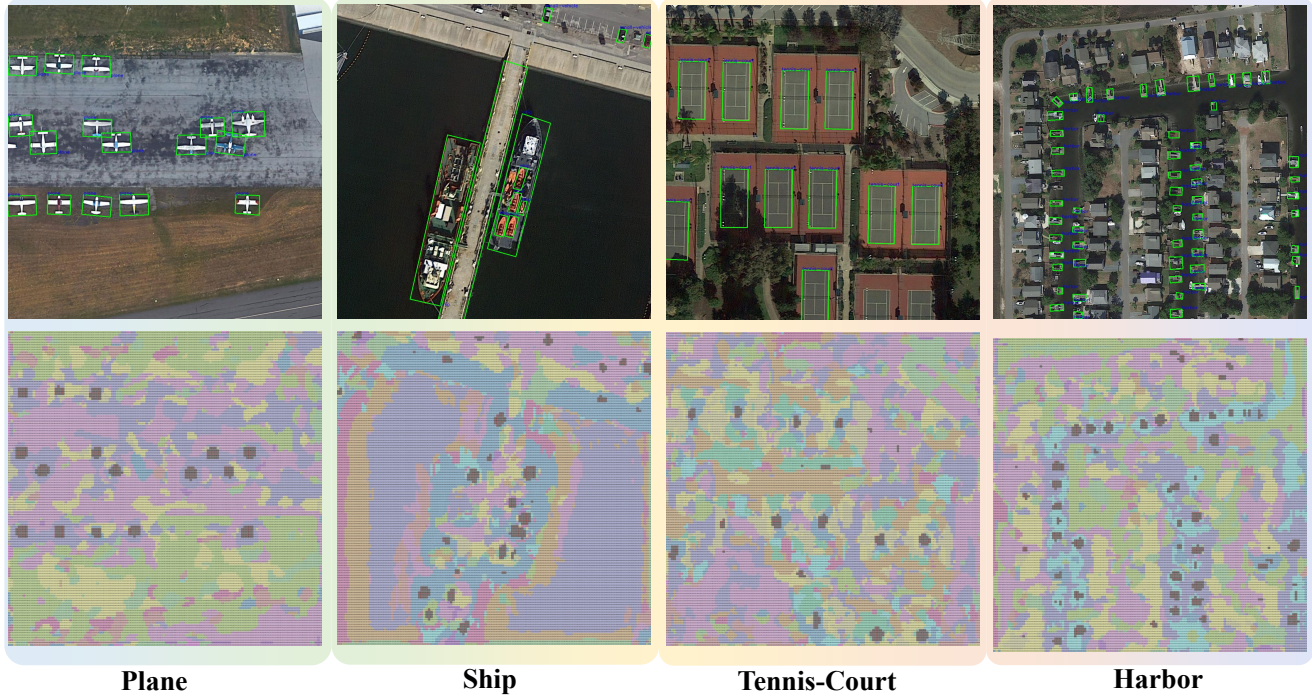


Figure 4. Illustration of visualized pixel selection results with the 5% labeling rate. The first row illustrates the original images, while the second row depicts the first-level feature maps extracted from the teacher logits. Gray dots denote the pixels identified as foreground.

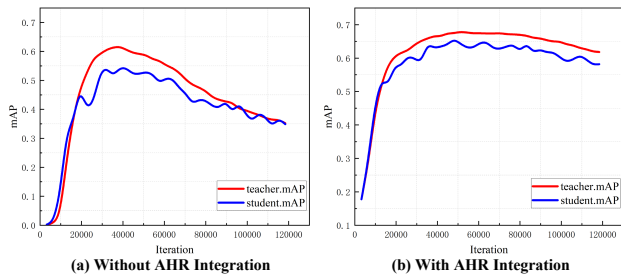


Figure 5. Impact of the AHR Integration on the mAP Performance.

mental influences. As depicted in Table 7, we conducted empirical assessments under both the “no-prompt” and “gt-prompt” configurations, which correspondingly delineate the upper-bound performance and facilitate systematic benchmarking against alternative assignment paradigms.

The impact of hyperparameter selection within AHR is systematically analyzed, with the corresponding experimental results summarized in Table 8. In this study, we investigate the effects of two key hyperparameters: the threshold parameter  $thr$ , which determines the selection criterion for hard negative samples, and the weighting factor  $w$ , which governs the degree of down-weighting. Our empirical findings indicate that while the overall performance remains relatively stable across different hyperparameter configurations, the optimal detection performance of 68.20% on the training set is achieved when the threshold  $thr$  is set to 0.9 and the weight  $w$  to 0.15. Moreover, Fig. 5 presents the comparative analysis of mAP curves before and after

applying AHR. A noticeable improvement in the stability of mAP progression is observed, wherein the curve exhibits a smoother convergence trend, eliminating the abrupt declines that are typically present in the absence of AHR. This empirically substantiates the effectiveness of our strategy in mitigating instability caused by noisy negative samples, ultimately leading to more reliable and robust learning dynamics throughout the training process.

## 5. Conclusion

This paper introduces an LLM-assisted semantic guidance framework based on a Multi-Branch Input architecture to enhance pseudo-label reliability for sparsely annotated object detection in remote sensing imagery. The Adaptive Hard-Negative Reweighting module alleviates optimization biases in the supervised branch, while the LLM-Assisted Semantic Prediction module refines class-specific prompts to facilitate Class-Aware Label Assignment. Extensive experiments on public benchmarks demonstrate the effectiveness of our approach in mitigating category imbalance and assignment ambiguities, achieving superior performance over existing methods. Future work will explore the broader applicability and adaptation of our method to varying levels of label sparsity.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grants Nos. 62372238, 62476133).

## References

- [1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in Neural Information Processing Systems*, 27, 2014. [2](#)
- [2] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14381–14390, 2022. [1](#)
- [3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. [6](#)
- [4] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. 2023. [4](#)
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. [6](#)
- [6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. [1](#)
- [7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015. [2](#)
- [8] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021. [2](#)
- [9] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. [2](#)
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. [6](#)
- [11] Wei Hua, Ding kang Liang, Jingyu Li, Xiaolong Liu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Sood: Towards semi-supervised oriented object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15558–15567, 2023. [1](#)
- [12] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. [3](#)
- [13] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, International Conference on Machine Learning*, page 896. Atlanta, 2013. [2](#)
- [14] Gang Li, Xiang Li, Yujie Wang, Yichao Wu, Ding Liang, and Shanshan Zhang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *European Conference on Computer Vision*, pages 457–472, 2022. [1](#)
- [15] Gang Li, Xiang Li, Yujie Wang, Wu Yichao, Ding Liang, and Shanshan Zhang. Dtg-ssod: Dense teacher guidance for semi-supervised object detection. *Advances in Neural Information Processing Systems*, 35:8840–8852, 2022. [2](#)
- [16] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. [5](#)
- [17] Ding kang Liang, Wei Hua, Chunsheng Shi, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Sood++: Leveraging unlabeled data to boost oriented object detection. *arXiv preprint arXiv:2407.01016*, 2024. [1](#)
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. [1](#)
- [19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2980–2988, 2017. [5](#)
- [20] Chang Liu, Weiming Zhang, Xiangru Lin, Wei Zhang, Xiao Tan, Junyu Han, Xiaomao Li, Errui Ding, and Jingdong Wang. Ambiguity-resistant semi-supervised learning for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15579–15588, 2023. [2](#), [3](#), [5](#)
- [21] Liang Liu, Bosheng Zhang, Jiangning Zhang, Wuhao Zhang, Zhenye Gan, Guanzhong Tian, Wenbing Zhu, Yabiao Wang, and Chengjie Wang. Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7370–7379, 2023. [3](#)
- [22] 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. In *International Conference on Learning Representations*, 2021. [2](#), [3](#)
- [23] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. [2](#), [3](#)
- [24] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International Conference on Pattern Recognition Applications and Methods*, pages 324–331, 2017. [2](#), [5](#)
- [25] Zihan Lu, Chenxu Wang, Chunyan Xu, Xiangwei Zheng, and Zhen Cui. Progressive exploration-conformal learn-

- ing for sparsely annotated object detection in aerial images. *Advances in Neural Information Processing Systems*, 37: 40593–40614, 2024. 1, 3, 6
- [26] Yusuke Niitani, Takuya Akiba, Tommi Kerola, Toru Ogawa, Shotaro Sano, and Shuji Suzuki. Sampling techniques for large-scale object detection from sparsely annotated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3
- [27] 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 *International Conference on Machine Learning*, pages 8748–8763, 2021. 4
- [28] Sai Saketh Rambhatla, Saksham Suri, Rama Chellappa, and Abhinav Shrivastava. Sparsely annotated object detection: A region-based semi-supervised approach. *arXiv preprint arXiv:2201.04620*, 2022. 1, 3
- [29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 2
- [30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016. 2
- [31] Saksham Suri, Saketh Rambhatla, Rama Chellappa, and Abhinav Shrivastava. Sparsedet: Improving sparsely annotated object detection with pseudo-positive mining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6770–6781, 2023. 1
- [32] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2291–2301, 2021. 1
- [33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2, 6
- [34] Chenxu Wang, Chunyan Xu, Xiang Li, YuXuan Li, Xu Guo, Ziqi Gu, and Zhen Cui. Multi-clue consistency learning to bridge gaps between general and oriented object in semi-supervised detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7582–7590, 2025. 2, 3
- [35] Haohan Wang, Liang Liu, Boshen Zhang, Jiangning Zhang, Wuhao Zhang, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Haoqian Wang. Calibrated teacher for sparsely annotated object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2519–2527, 2023. 3
- [36] Tiancai Wang, Tong Yang, Jiale Cao, and Xiangyu Zhang. Co-mining: Self-supervised learning for sparsely annotated object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2800–2808, 2021. 3
- [37] Xinjiang Wang, Xingyi Yang, Shilong Zhang, Yijiang Li, Litong Feng, Shijie Fang, Chengqi Lyu, Kai Chen, and Wayne Zhang. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3240–3249, 2023. 1
- [38] Wenhao Wu, Hau-San Wong, and Si Wu. Pseudo-siamese teacher for semi-supervised oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3
- [39] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 2, 5
- [40] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3520–3529, 2021. 2
- [41] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 2, 3
- [42] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *European Conference on Computer Vision*, pages 677–694, 2020. 2
- [43] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3163–3171, 2021. 2
- [44] Fangyuan Zhang, Tianxiang Pan, and Bin Wang. Semi-supervised object detection with adaptive class-rebalancing self-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3252–3261, 2022. 3
- [45] Han Zhang, Fangyi Chen, Zhiqiang Shen, Qiqi Hao, Chenchen Zhu, and Marios Savvides. Solving missing-annotation object detection with background recalibration loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1888–1892, 2020. 5
- [46] Wenqing Zhao and Lijiao Xu. Weakly supervised target detection based on spatial attention. *Visual Intelligence*, 2(1): 2, 2024. 2
- [47] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *European Conference on Computer Vision*, pages 35–50, 2022. 2, 3, 6, 7
- [48] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, et al. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7331–7334, 2022. 6