SEMANTIC PRIOR-GUIDED LEARNING WITH ADAPTIVE PSEUDO-LABELING FOR SEMI-SUPERVISED ORIENTED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-supervised oriented object detection provides a promising solution for reducing annotation costs in remote-sensing imagery. However, existing approaches still face two critical challenges: insufficient representation capability caused by limited labeled data and unreliable pseudo-labels generated from unlabeled data. These challenges are further exacerbated by arbitrary object orientations, dense object distributions, substantial scale variations, and complex inter-class similarities in remote-sensing imagery. To address these issues, we propose SPAL-Net, a semantic prior-guided learning framework with adaptive pseudo-labeling for semi-supervised oriented object detection. SPAL-Net leverages CLIP-derived text embeddings as class-level semantic priors to enhance visual representation learning. Specifically, a Semantic Alignment Module (SAM) is designed to project multi-scale visual features into the text embedding space and enhance visual representations with class-aware semantic information. Furthermore, an Instance-Level Distribution Alignment Loss (IDAL) is proposed to provide semantic supervision by aligning the visual class distribution of each positive instance with the corresponding text-induced class distribution. This enables the model to exploit inter-class semantic relationships beyond conventional one-hot supervision, thereby improving the discriminative capability of instance representations. To improve the quality of unlabeled supervision, we develop a Class-Adaptive Pseudo-Label Filtering (CAPF) strategy. By maintaining class-specific confidence thresholds using exponential moving averages, CAPF dynamically adjusts pseudo-label selection based on evolving class-wise confidence statistics, effectively filtering out noisy pseudo-labels and improving the reliability of unlabeled supervision. Extensive experiments on benchmark remote-sensing datasets demonstrate that SPAL-Net achieves state-of-the-art performance under various partial-label settings, validating its effectiveness.

1 INTRODUCTION

Oriented object detection is fundamental to remote-sensing image analysis, enabling accurate localization of arbitrarily oriented objects in complex aerial scenes (Xia et al., 2018; Cheng et al., 2022). However, current methods depend on extensive high-quality annotations, whose acquisition is costly. Semi-supervised oriented object detection (SSOOD) addresses this annotation bottleneck by jointly training a detector on a small labeled set and a large unlabeled set, thereby reducing annotation costs while improving generalization. Although semi-supervised methods have narrowed the performance gap in horizontal object detection (Liu et al., 2021; Xu et al., 2021; Zhou et al., 2022a), transferring them to SSOOD in remote-sensing imagery remains challenging because these scenes contain arbitrarily oriented, densely distributed objects with substantial scale variation, making reliable pseudo-label supervision particularly important. Most existing methods adopt an EMA teacher–student framework (Tarvainen & Valpola, 2017). Within this paradigm, two limitations remain. First, limited labeled data restrict visual representation learning, while one-hot supervision discards latent semantic relationships among classes. Second, class imbalance and heterogeneous detection difficulty produce confidence distributions that vary across classes and training stages, making a single fixed threshold unsuitable for pseudo-label selection.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To strengthen representation learning, vision–language models such as CLIP (Radford et al., 2021) provide a means of introducing textual priors. Class text embeddings encode semantic relationships learned from large-scale image–text pretraining and can complement the one-hot supervision available from limited labeled data. However, integrating these embeddings into SSOOD raises three challenges: class-level text embeddings are structurally different from multi-scale spatial features; conventional RoI classification does not constrain visual class distributions to reflect text-induced class relationships; and incorrect pseudo-label class assignments may associate unlabeled instances with mismatched text-induced distributions.

To address these issues, we propose SPAL-Net, a semantic prior-guided learning framework with adaptive pseudo-labeling for SSOOD. SAM maps multi-scale visual features into the CLIP text-embedding space and uses gated attention to inject class context into FPN features. IDAL aligns each RoI's visual class distribution with a text-induced distribution derived from similarities among class embeddings, allowing inter-class semantics to complement one-hot supervision. CAPF tracks class-wise confidence and adapts filtering thresholds during training to reduce noise from evolving confidence distributions.

In summary, our main contributions are as follows:

- We propose SPAL-Net, which jointly integrates textual priors into feature-level and instance-level learning and adapts pseudo-label filtering to class-wise confidence dynamics for semi-supervised oriented object detection, thereby improving semantic representation learning and the reliability of unlabeled supervision.
- We design a Semantic Alignment Module that transforms global class embeddings into spatially resolved semantic context through dense visual–text response aggregation and gated modulation of multi-scale FPN features, enabling class-aware semantic information to enhance features at relevant locations and scales.
- We develop an Instance-Level Distribution Alignment Loss that derives text-induced class distributions from the self-similarity of class embeddings and aligns them with RoI visual class distributions, allowing inter-class semantic relationships to complement one-hot supervision and improve instance-level class discrimination.
- We introduce Class-Adaptive Pseudo-Label Filtering, which replaces a single fixed threshold with class-specific thresholds updated from evolving teacher confidence statistics, helping reduce noisy pseudo-labels under class imbalance and changing detection difficulty.

2 Related Work

Semi-Supervised Oriented Object Detection. Mean Teacher establishes the basic paradigm of modern semi-supervised object detection by using an EMA-updated teacher to provide stable learning targets for the student (Tarvainen & Valpola, 2017). Building on this paradigm, Unbiased Teacher reduces pseudo-label bias through confidence filtering and focal loss (Liu et al., 2021), Soft Teacher weights unlabeled classification samples using teacher confidence (Xu et al., 2021), Dense Teacher introduces dense pseudo-label supervision (Zhou et al., 2022a), and PseCo combines pseudo-labeling with consistency training across data views (Li et al., 2022a).

For oriented object detection, SOOD addresses orientation-related pseudo-label inconsistency in remote-sensing imagery (Hua et al., 2023). Subsequent methods improve dense prediction and pseudo-label reliability through ambiguity-resistant learning (Liu et al., 2023), pseudo-Siamese teacher networks (Wu et al., 2024), density-guided pseudo-label selection (Zhao et al., 2024), global focal learning (Wang et al., 2024a), denser oriented pseudo-label supervision (Zhao et al., 2025a), and joint representation–supervision enhancement in SROD (Meng et al., 2026). Oriented Teacher (Li et al., 2026) exploits spatial-resolution disparity between teacher and student views, while SOOD++ (Liang et al., 2026) further enhances pseudo-label reliability through informative instance selection and geometric quality modeling. Overall, existing SSOOD methods focus mainly on the visual and geometric aspects of pseudo-label supervision. In contrast, SPAL-Net uses CAPF for class-adaptive pseudo-label filtering and further incorporates the semantic relationships encoded by class text embeddings into multi-scale features and RoI-level supervision.

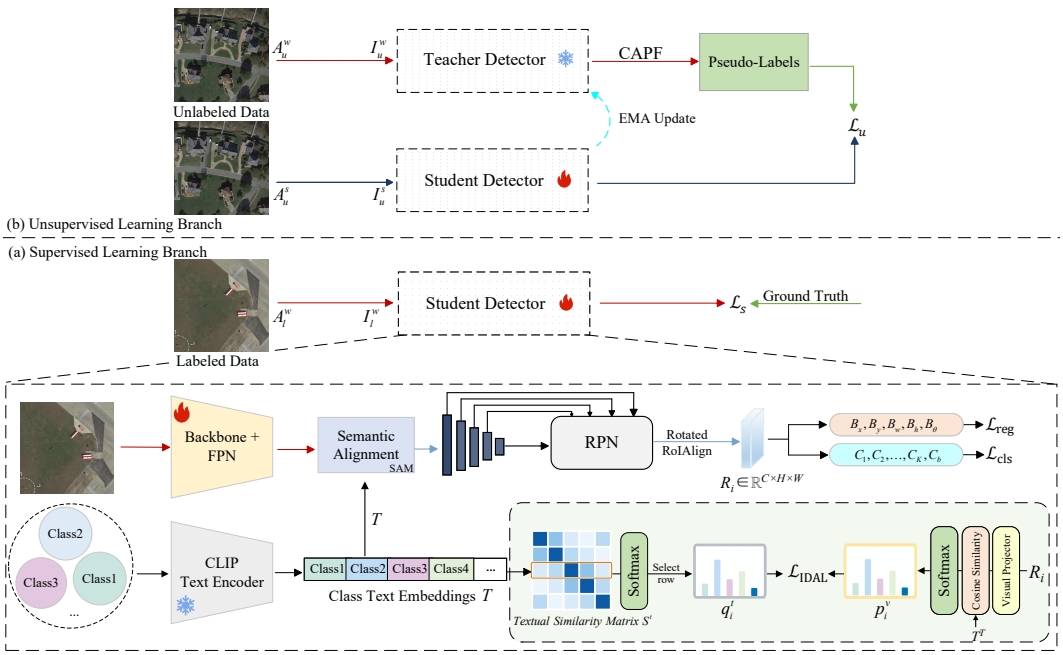


Figure 1: Overview of SPAL-Net. The supervised and unsupervised branches share the student detector; CAPF filters teacher predictions to form pseudo-labels, and the teacher is updated by EMA. SAM injects frozen CLIP-derived textual priors into FPN features, while IDAL aligns RoI visual class distributions with text-induced class distributions.

Textual Semantic Priors for Object Detection. Vision–language pretraining introduces transferable semantic knowledge into object detection. CLIP (Radford et al., 2021) learns a shared image–text embedding space through large-scale contrastive learning, enabling open-vocabulary detectors to use text for region classification and localization. Representative approaches to open-vocabulary object detection include ViLD (Gu et al., 2022) for vision–language distillation; RegionCLIP (Zhong et al., 2022), GLIP (Li et al., 2022b), and DetCLIPv2 (Yao et al., 2023) for region- or word-level alignment; DetPro (Du et al., 2022) for prompt learning; and DetCLIP (Yao et al., 2022) for detection-oriented visual-concept pretraining. F-VLM (Kuo et al., 2023) further validates frozen vision–language representations for region-level recognition, showing that text embeddings can serve as classifiers, alignment targets, and semantic priors for object detection.

Textual priors have also been explored in remote-sensing image understanding. RemoteCLIP (Liu et al., 2024) and GeoRSCLIP (Zhang et al., 2024) learn domain-specific image–text semantics, CastDet (Li et al., 2024) and LAE (Pan et al., 2025) transfer such priors to open-vocabulary aerial detection, and Zhou et al. (2025) align oriented RoI features with CLIP text embeddings for fully supervised fine-grained oriented detection. However, how to coordinate textual priors across dense feature learning and instance-level supervision remains underexplored. SPAL-Net addresses this gap at two complementary levels: SAM enriches multi-scale FPN features with class-aware context, while IDAL uses relationships among frozen class embeddings to define text-induced class distributions for RoI-level supervision.

3 Method

3.1 Overview

As shown in Figure 1, SPAL-Net builds on the Mean Teacher paradigm (Tarvainen & Valpola, 2017) with Rotated Faster R-CNN (Ren et al., 2015). The student is optimized by backpropagation, and the teacher is updated by an exponential moving average (EMA) of the student. Both use a shared detection path with SAM-enhanced backbone/FPN features passed to the RPN and RoI heads, with frozen CLIP text embeddings injected before proposal generation. For labeled images, the student

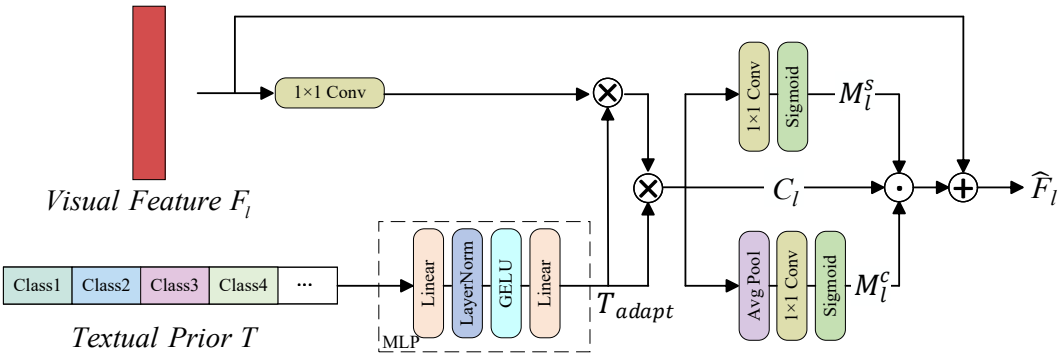


Figure 2: Architecture of the Semantic Alignment Module (SAM), which injects class-aware textual priors into multi-scale FPN features.

uses conventional detection losses and RoI-level IDAL based on ground-truth labels. For unlabeled images, the teacher predicts classes, oriented boxes, and confidence scores from weakly augmented views; CAPF filters these predictions before pseudo-label construction. The student processes the corresponding strongly augmented views through the same detection path and computes IDAL using retained pseudo-labels.

3.2 Semantic Alignment Module

We construct a class-semantic prior by encoding class information with a frozen CLIP text encoder. Given the class set $\mathcal{C} = \{c_k\}_{k=1}^{K}$, the prompt for the k-th class is $p_k =$ "a photo of a c_k". The normalized class features and textual prior matrix are

$$\begin{aligned} t_k &= \frac{E_{\text{text}}(p_k)}{\|E_{\text{text}}(p_k)\|_2}, \\ T &= [t_1, t_2, \ldots, t_K]^\top \in \mathbb{R}^{K\times D}. \end{aligned} \tag{1}$$

Here, E_{text} is the frozen CLIP text encoder, and K and D denote the number of classes and the text-embedding dimension, respectively. The textual prior T is used for both FPN feature enhancement and subsequent RoI-level distribution alignment.

The textual prior T encodes class semantics as global class-level embeddings, whereas FPN features retain channel and spatial structure. To make these structurally different representations compatible for dense visual–text interaction, SAM projects the FPN features into the text embedding space and adapts the textual prior. Figure 2 summarizes the subsequent response aggregation, gated modulation, and residual update. For the l-th FPN feature $F_l \in \mathbb{R}^{B\times C\times H_l\times W_l}$, a 1×1 visual projection produces $X_l = \text{Conv}_{1\times1}(F_l) \in \mathbb{R}^{B\times D\times H_l\times W_l}$. Meanwhile, the textual prior is normalized and adapted as $T_{\text{adapt}} = \text{MLP}(\text{LayerNorm}(T))$. The text adapter uses two linear layers with an intermediate GELU and is optimized jointly with the detector to reduce the remote-sensing domain gap.

We flatten X_l into $X_l^\flat \in \mathbb{R}^{B\times D\times N_l}$, where $N_l = H_l W_l$. We then compute the scaled dot-product similarity between the visual feature at each spatial location and every adapted class embedding, followed by a softmax over the class dimension:

$$R_l = \text{Softmax}_K\left(\frac{T_{\text{adapt}} X_l^\flat}{\sqrt{D}}\right), \qquad R_l \in \mathbb{R}^{B\times K\times N_l}. \tag{2}$$

R_l represents dense class responses, enabling each spatial location to select class semantics relevant to its visual content. Aggregating the adapted text features according to these responses yields a semantic context with spatial structure:

$$C_l = \text{Reshape}\left(T_{\text{adapt}}^\top R_l\right), \qquad C_l \in \mathbb{R}^{B\times D\times H_l\times W_l}. \tag{3}$$

To suppress textual responses irrelevant to the current image content, SAM modulates the semantic context along the spatial and channel dimensions:

$$\begin{aligned} M_l^s &= \sigma\left(\text{Conv}_{1\times1}(C_l)\right), \\ M_l^c &= \sigma\left(\text{Conv}_{1\times1}(\text{GAP}(C_l))\right), \end{aligned} \tag{4}$$

where GAP denotes global average pooling and σ is the sigmoid function. The spatial gate $M_l^s \in \mathbb{R}^{B\times 1\times H_l\times W_l}$ selects locations that require semantic enhancement, while the channel gate $M_l^c \in \mathbb{R}^{B\times D\times 1\times 1}$ controls the contribution of different semantic channels. The enhanced FPN feature is

$$\widehat{F}_l = F_l + \gamma \mathrm{Conv}_{1\times 1}\left(C_l \odot M_l^s \odot M_l^c\right), \tag{5}$$

where the 1×1 convolution projects the semantic context back to the FPN channel space, $\odot$ denotes element-wise multiplication, and γ is a learnable scaling factor initialized to 2×10^{-4}. All $\mathrm{Conv}_{1\times 1}$ operations in SAM denote separate convolutional mappings with independently learned parameters. This residual design preserves the original visual representation early in training while allowing the strength of textual semantic injection to be learned gradually.

SAM parameters are shared across FPN levels. SAM enhances pre-proposal multi-scale features in both student training branches, during teacher pseudo-label generation, and at inference.

3.3 Instance-Level Distribution Alignment Loss

SAM introduces class-semantic context at the FPN level, but one-hot RoI supervision cannot explicitly model inter-class semantic relationships. We therefore propose text-guided Instance-Level Distribution Alignment (IDAL), which derives a text-induced class distribution for each class from the self-similarity of CLIP text features and aligns each RoI's visual class distribution with the corresponding text-induced distribution.

Given the i-th instance feature $R_i \in \mathbb{R}^{C\times H\times W}$ output by RoIAlign, we map it into the CLIP text embedding space and compute its visual class distribution using the textual prior T:

$$\begin{aligned} v_i &= \mathrm{Norm}\left(g_v\left(\mathrm{GAP}(R_i)\right)\right), \\ p_i^v &= \mathrm{Softmax}\left(\frac{v_i T^\top}{\tau_v}\right). \end{aligned} \tag{6}$$

Here, g_v consists of two fully connected layers with an intermediate ReLU activation, Norm denotes L2 normalization, and τ_v controls the smoothness of the visual class distribution.

Because the class text features in T are normalized, their self-similarity matrix and the resulting matrix of text-induced class distributions are

$$\begin{aligned} S^t &= TT^\top, \\ Q^t &= \mathrm{Softmax}\left(\frac{S^t}{\tau_t}\right). \end{aligned} \tag{7}$$

Here, $S^t_{a,b}$ denotes the semantic similarity between classes c_a and c_b, and τ_t is the textual temperature. For a positive RoI with class label y_i, the corresponding text-induced class distribution is $q_i^t = Q^t_{y_i,:}$. Unlike one-hot labels, q_i^t preserves the dominant signal of the assigned class while assigning nonzero probabilities to semantically related classes. The distribution is computed from the frozen textual prior and receives no gradients.

IDAL minimizes the KL divergence from each text-induced class distribution to the corresponding visual class distribution:

$$\mathcal{L}_{\mathrm{IDAL}} = \frac{1}{N_+}\sum_{i=1}^{N_+} D_{\mathrm{KL}}\left(q_i^t \,\|\, p_i^v\right), \tag{8}$$

where N_+ denotes the number of positive RoIs; background and invalid RoIs are excluded. The temperatures τ_v and τ_t control the smoothness of the visual and text-induced class distributions, respectively.

We compute $\mathcal{L}_{\mathrm{IDAL}}$ at the RoI stage of both student training branches. In the supervised branch, positive RoIs are assigned ground-truth class labels, which are used to select the corresponding text-induced class distributions q_i^t. The resulting IDAL loss is combined with the standard RPN and RoI detection losses to form the supervised branch objective $\mathcal{L}_s$. In the unsupervised branch, CAPF-retained pseudo-labels are used to assign positive RoIs and select their corresponding q_i^t. The IDAL loss computed from these RoIs is combined with the standard unsupervised RPN and RoI detection losses to form $\mathcal{L}_u$. In both branches, $\mathcal{L}_{\mathrm{IDAL}}$ is incorporated with unit weight. The classification loss enforces class discrimination, whereas minimizing $\mathcal{L}_{\mathrm{IDAL}}$ encourages the visual class distributions to reflect text-encoded inter-class relationships.

Table 1: Comparison on DOTA-v1.5 and DOTA-v1.0 under the partial-label protocol. All results are reported in mAP (%). The best and second-best results for each dataset are shown in bold and underlined, respectively; – indicates that a compatible result is unavailable. Results for Oriented Teacher and SROD are taken from their original publications; Oriented Teacher uses Faster R-CNN OBB/Oriented R-CNN, whereas SROD uses Rotated FCOS.

		DOTA-v1.5			DOTA-v1.0		
Setting	Method	10%	20%	30%	10%	20%	30%
Supervised	Faster R-CNN* (Ren et al., 2015)	43.43	51.32	53.14	44.96	55.98	57.89
	FCOS† (Tian et al., 2019)	42.78	50.11	54.79	44.83	54.16	57.41
Semi-supervised	Unbiased Teacher* (Liu et al., 2021)	44.51	52.80	53.33	–	–	–
	Soft Teacher* (Xu et al., 2021)	48.46	54.89	57.86	–	–	–
	Dense Teacher† (Zhou et al., 2022a)	46.90	53.93	57.86	50.87	58.02	60.91
	PseCo* (Li et al., 2022a)	48.04	55.28	59.03	–	–	–
	DualPolish* (Zhang et al., 2023)	49.02	55.17	58.44	–	–	–
	SOOD† (Hua et al., 2023)	48.63	55.58	59.23	52.11	59.23	61.44
	ARSL† (Liu et al., 2023)	47.56	55.05	58.94	52.89	59.84	62.92
	PST* (Wu et al., 2024)	49.63	57.39	60.40	–	–	–
	DDPLS† (Zhao et al., 2024)	51.39	57.12	59.00	–	–	–
	Active Teacher† (Wang et al., 2024b)	50.57	57.13	61.04	–	–	–
	Focal Teacher† (Wang et al., 2024a)	52.24	57.00	60.21	–	–	–
	Denser Teacher† (Zhao et al., 2025a)	52.05	57.49	60.40	52.62	59.20	62.82
	TMR-RD-v2* (Marvasti-Zadeh et al., 2024)	47.88	55.17	59.02	52.81	59.90	63.12
	SRODET† (Wang et al., 2025)	49.23	55.38	59.16	–	–	–
	Adjacent Teacher† (Xia et al., 2025)	51.26	57.42	60.26	–	–	–
	SAOD† (Zhu et al., 2025)	51.01	57.14	60.44	–	–	–
	DCPLS† (Yang et al., 2025)	49.23	55.71	59.10	–	–	–
	CBLC-SOOD† (Zhao et al., 2025b)	49.64	56.12	59.36	52.77	58.70	62.79
	SOOD++† (Liang et al., 2026)	50.48	57.44	<u>61.51</u>	<u>54.17</u>	<u>60.53</u>	<u>64.93</u>
	Oriented Teacher* (Li et al., 2026)	<u>52.30</u>	<u>57.76</u>	61.23	–	–	–
	SROD† (Meng et al., 2026)	50.54	57.60	60.95	52.58	55.12	59.41
	SPAL-Net* (Ours)	**54.77**	**60.72**	**62.84**	**58.97**	**63.38**	**66.93**

Combining the supervised and unsupervised branches, the overall training objective of SPAL-Net is

$$\mathcal{L} = \mathcal{L}_s + \lambda_u(t)\mathcal{L}_u, \tag{9}$$

where $\mathcal{L}_s$ and $\mathcal{L}_u$ denote the complete supervised and unsupervised branch objectives, respectively, each including its corresponding IDAL term. Here, t denotes the current training iteration, and $\lambda_u(t)$ is an iteration-dependent weight that controls the contribution of $\mathcal{L}_u$ to the overall training objective. The weight $\lambda_u(t)$ is set to zero before the unsupervised branch is activated, increases linearly from zero to one during the initial ramp-up interval following its activation, and remains one thereafter.

3.4 Class-Adaptive Pseudo-Label Filtering

Fixed-threshold selection assumes confidence scores are comparable across classes and stable during training, an assumption violated by class-frequency, object-scale, and difficulty differences in remote-sensing oriented detection. As class-wise confidence distributions evolve, a high threshold can reject valid predictions from difficult classes, whereas a low threshold can admit noise. CAPF therefore applies a global pre-filter, then updates per-class thresholds from current confidence statistics to balance shared quality control with class-specific adaptation.

At training iteration r, given the weakly augmented unlabeled input I_u^w, the teacher produces M candidate tuples $(\hat{b}_i, \hat{c}_i, s_i)$ after detector post-processing, where $\hat{b}_i$, $\hat{c}_i$, and s_i denote the oriented bounding box, class label, and classification confidence, respectively. CAPF first applies a global threshold δ to remove low-confidence predictions. For each class c, the surviving indices are $\mathcal{I}_c^{(r)} =$

Table 2: Comparison under the full-label semi-supervised protocol. Each entry reports the supervised detector baseline → the semi-supervised result in mAP (%). The best and second-best semi-supervised results are shown in bold and underlined, respectively. Results for Oriented Teacher are taken from its original publication and use Faster R-CNN OBB/Oriented R-CNN; – indicates that a compatible result is unavailable.

Method	DOTA-v1.5	DOTA-v1.0
Unbiased Teacher* (Liu et al., 2021)	66.12 → 64.85	–
Soft Teacher* (Xu et al., 2021)	66.12 → 66.40	–
Dense Teacher† (Zhou et al., 2022a)	65.46 → 66.38	70.22 → 71.20
ARSL† (Liu et al., 2023)	65.46 → 66.68	70.22 → 71.63
SOOD† (Hua et al., 2023)	65.46 → 67.70	–
DDPLS† (Zhao et al., 2024)	65.46 → 66.93	–
Focal Teacher† (Wang et al., 2024a)	65.46 → 67.12	–
TMR-RD-v2* (Marvasti-Zadeh et al., 2024)	66.12 → 66.93	70.08 → 71.68
SRODET† (Wang et al., 2025)	65.34 → 67.12	–
Adjacent Teacher† (Xia et al., 2025)	66.23 → 67.39	–
AODet† (Chen et al., 2025)	65.46 → 67.78	–
SAOD† (Zhu et al., 2025)	65.46 → 67.82	–
CBLC-SOOD† (Zhao et al., 2025b)	65.46 → 66.61	70.22 → 70.67
SOOD++† (Liang et al., 2026)	65.46 → **69.04**	70.22 → <u>72.40</u>
Oriented Teacher* (Li et al., 2026)	66.12 → 68.05	–
SPAL-Net* (Ours)	66.12 → <u>68.55</u>	70.08 → **72.59**

$\{i \mid \hat{c}_i = c,\ s_i > \delta\}$. When $\mathcal{I}_c^{(r)} \neq \emptyset$, the mean confidence of class c at the current iteration is

$$\bar{s}_c^{(r)} = \frac{1}{\left|\mathcal{I}_c^{(r)}\right|} \sum_{i \in \mathcal{I}_c^{(r)}} s_i. \tag{10}$$

CAPF uses this statistic to maintain a dynamic threshold for each observed class. The threshold is initialized by the current mean confidence when class c first appears and is subsequently updated by an exponential moving average:

$$\tau_c^{(r)} = \begin{cases} \bar{s}_c^{(r)}, & \text{if class } c \text{ is first observed,} \\ \rho \tau_c^{(r-1)} + (1-\rho)\bar{s}_c^{(r)}, & \text{otherwise,} \end{cases} \tag{11}$$

where ρ is the threshold momentum. The EMA update smooths class-wise confidence fluctuations. If no prediction survives for an initialized class at iteration r, its threshold remains unchanged; unseen classes remain inactive until they are first observed.

After updating the class-specific thresholds, CAPF applies a second filtering stage and constructs the final pseudo-label set as

$$\widehat{\mathcal{Y}}_u^{(r)} = \left\{ (\hat{b}_i, \hat{c}_i) \;\middle|\; i \in \bigcup_{c=1}^{K} \mathcal{I}_c^{(r)},\ s_i > \tau_{\hat{c}_i}^{(r)} \right\}. \tag{12}$$

4 Experiments

Datasets. We conduct experiments on DOTA-v1.0 and DOTA-v1.5 under different annotation budgets. DOTA-v1.0 (Xia et al., 2018) contains 2,806 aerial images, including 1,411 training images, 458 validation images, and 937 test images, with 188,282 oriented bounding boxes across 15 classes. DOTA-v1.5 uses the same image split, additionally annotates objects smaller than 10 pixels, and introduces the container crane class, resulting in 16 classes and 403,318 annotated instances.

We use two semi-supervised evaluation protocols. 1) *Partial-label setting*: Following the fixed splits of previous work (Hua et al., 2023), we use 10%, 20%, or 30% of the training images as labeled

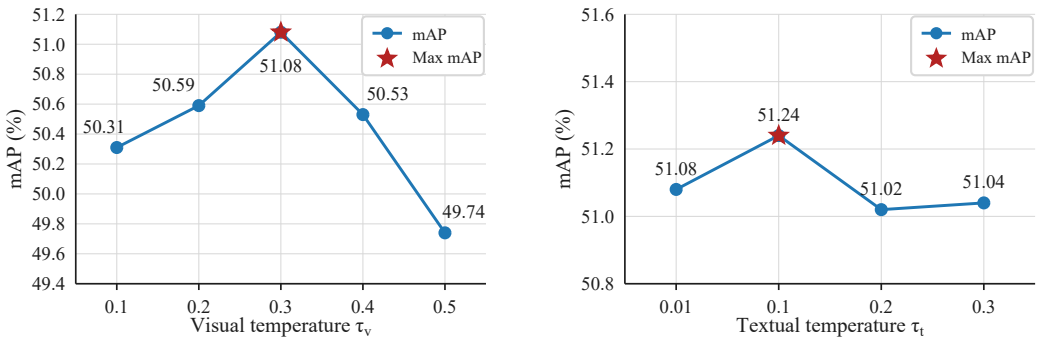


Figure 3: Temperature sensitivity of IDAL on DOTA-v1.5 under the 10% annotation protocol. Left: visual temperature τ_v. Right: textual temperature τ_t. Red stars indicate the best result in each sweep.

data and the remainder as unlabeled data, and evaluate all models on the validation set. 2) *Full-label setting*: We use the complete training set as labeled data, the corresponding test set as unlabeled data, and the validation set for evaluation. Because training still uses additional unlabeled images, we refer to this protocol as the full-label semi-supervised setting. We follow the standard evaluation protocol of each dataset and report mean average precision (mAP) (Xia et al., 2018). mAP is the standard evaluation metric for DOTA, which facilitates fair comparison with prior work.

Implementation Details. We implement SPAL-Net in PyTorch (Paszke et al., 2019) with MMRotate (Zhou et al., 2022b). We use Rotated Faster R-CNN (Ren et al., 2015) with an FPN (Lin et al., 2017) and an ImageNet-pretrained ResNet-50 backbone. Class text embeddings are extracted by a pretrained CLIP ViT-B/16 text encoder (Radford et al., 2021), which remains frozen throughout training. All experiments run on a single NVIDIA GeForce RTX 3090 GPU, with each batch containing two labeled images and one unlabeled image. We optimize the model using SGD with a momentum of 0.9 and a weight decay of 10^{-4}. DOTA images are cropped into 1024×1024 patches with a 200-pixel overlap and an 824-pixel stride.

For unlabeled data, the teacher receives weakly augmented views generated by random flipping and rotation, while the student receives strongly augmented views with additional color jittering, random grayscale conversion, Gaussian blurring, and random cropping. EMA updates for the teacher begin after 3,200 iterations with a momentum of 0.9996, and the unsupervised branch is activated after a burn-in interval of 6,400 iterations, followed by a ramp-up interval of 6,400 iterations. We train all DOTA partial-label configurations for 180k iterations, with initial learning rates ranging from 6.25×10^{-4} to 5×10^{-3}. Across configurations, the CAPF global threshold ranges from 0.6 to 0.8, and the class-specific threshold updates use an independent EMA momentum of 0.9996. The IDAL visual and textual temperatures are set to 0.3 and 0.1, respectively.

4.1 Comparison with State-of-the-Art Methods

To validate the effectiveness of SPAL-Net, we compare it with previous methods under the DOTA-v1.5-Partial, DOTA-v1.0-Partial, DOTA-v1.5-Full, and DOTA-v1.0-Full protocols. Oriented implementations of several general SSOD methods are also included. In the tables, * and † denote Rotated Faster R-CNN (Ren et al., 2015) and Rotated FCOS (Tian et al., 2019), respectively.

Partial-label setting. Table 1 presents the results under the partial-label protocols. On DOTA-v1.5, SPAL-Net achieves 54.77, 60.72, and 62.84 mAP with 10%, 20%, and 30% labeled data, respectively, yielding absolute gains of 11.34, 9.40, and 9.70 mAP points over the corresponding supervised Rotated Faster R-CNN baselines. Compared with the strongest previous results, reported by Oriented Teacher (Li et al., 2026) at 10% and 20% labeling and by SOOD++ (Liang et al., 2026) at 30% labeling, SPAL-Net achieves further gains of 2.47, 2.96, and 1.33 mAP points, respectively. On DOTA-v1.0, SPAL-Net obtains 58.97, 63.38, and 66.93 mAP under the 10%, 20%, and 30% settings, respectively, corresponding to absolute gains of 14.01, 7.40, and 9.04 mAP points over the supervised baselines. Compared with the strongest previous results reported by the publicly

Table 3: Component ablation under the 10% annotation protocol. Results are reported in mAP (%).

CAPF	SAM	IDAL	DOTA-v1.0	DOTA-v1.5
–	–	–	49.16	45.09
✓	–	–	52.14	48.78
–	✓	–	55.40	51.83
–	–	✓	55.41	51.24
✓	✓	✓	**58.97**	**54.77**

available SOOD++ (Liang et al., 2026), SPAL-Net achieves further gains of 4.80, 2.85, and 2.00 mAP points. These results demonstrate that SPAL-Net consistently benefits from unlabeled data across both DOTA benchmarks and annotation budgets.

Full-label setting. Table 2 presents the results under the full-label semi-supervised protocol. On DOTA-v1.5, SPAL-Net achieves 68.55 mAP, improving the supervised baseline by 2.43 mAP points and ranking second, 0.49 mAP points below SOOD++ (Liang et al., 2026). On DOTA-v1.0, SPAL-Net reaches 72.59 mAP, outperforming the supervised baseline by 2.51 mAP points and surpassing SOOD++ by 0.19 mAP points. These results show that SPAL-Net achieves the best DOTA-v1.0 performance and remains competitive on DOTA-v1.5, while consistently benefiting from additional unlabeled images when labeled data are abundant.

4.2 Ablation Studies

We ablate CAPF, SAM, and IDAL under the 10% protocol on both datasets using the same Rotated Faster R-CNN teacher–student framework (Table 3). The baseline obtains 49.16/45.09 mAP on DOTA-v1.0/v1.5. CAPF alone adds 2.98/3.69 points, while SAM reaches 55.40/51.83 (+6.24/+6.74) and IDAL reaches 55.41/51.24 (+6.25/+6.15), validating both semantic components. SAM supplies dense class-aware context, while IDAL transfers text-encoded inter-class relationships to positive RoIs. CAPF separately strengthens unlabeled supervision by filtering class-wise predictions. Together, the three components achieve 58.97/54.77 mAP, gains of 9.81/9.68 over baseline and higher than any single-component variant, demonstrating their complementarity.

Temperature Sensitivity of IDAL. Figure 3 evaluates one temperature at a time, keeping other settings fixed, on DOTA-v1.5 with 10% labels. As τ_v increases from 0.1 to 0.3, mAP rises from 50.31 to 51.08, then falls to 50.53 at 0.4 and 49.74 at 0.5, favoring the intermediate value that balances concentration and smoothing. This trend suggests that lower temperatures over-concentrate visual predictions, whereas excessive smoothing at higher temperatures weakens class discrimination. For $\tau_t = 0.01, 0.10, 0.20, 0.30$, mAP is 51.08, 51.24, 51.02, and 51.04, a 0.22-point range. Thus IDAL is relatively robust to τ_t; moderate smoothing at 0.1 preserves class discrimination and semantic relationships and gives the best tested result. We use $\tau_v = 0.3$ and $\tau_t = 0.1$.

5 Conclusion

We presented SPAL-Net, a semantic prior-guided learning framework with adaptive pseudo-labeling for semi-supervised oriented object detection. By dynamically filtering noisy pseudo-labels and using textual priors to enhance multi-scale features and regularize instance-level distributions, SPAL-Net improves the reliability of unlabeled supervision and the semantic discriminability of visual representations. Experiments on DOTA-v1.0 and DOTA-v1.5 validate its effectiveness across annotation budgets and DOTA versions. However, the method still has several limitations: CAPF does not explicitly consider localization quality, angle uncertainty, or object density; CLIP text embeddings generated from fixed class names may not capture fine-grained visual differences; and incorrect pseudo-label class assignments may introduce misleading supervision into IDAL. Future work will further model the geometric quality of pseudo-labels, explore more adaptive vision–language priors for remote-sensing imagery, and improve the robustness of IDAL supervision to noisy pseudo-labels.

AI USE STATEMENT

In preparing this manuscript, the authors used generative AI tools for language editing and polishing. The authors did not use generative AI tools to generate research ideas, formulate or prove mathematical claims, design or conduct experiments, generate or clean datasets, implement the proposed method, or interpret experimental results. All AI-assisted outputs were reviewed and verified by the authors, including the manuscript text, citations, technical claims, equations, and experimental descriptions. The authors take full responsibility for the final content of this work.

REFERENCES

Jiecong Chen, Chenlin Fu, and Yingying Zhu. Adaptive optimization strategy for semi-supervised arbitrary-oriented object detection. In *2025 IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11210067.

Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. doi: 10.1109/TGRS.2022.3183022.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14064–14073, 2022. doi: 10.1109/CVPR52688.2022.01369.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2104.13921.

Wei Hua, Dingkang 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*, pp. 15558–15567, 2023. doi: 10.1109/CVPR52729.2023.01493.

Weicheng Kuo, Yin Cui, Xiuye Gu, A. J. Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MIMwy4kh9lf.

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 *Computer Vision – ECCV 2022*, pp. 457–472. Springer Nature Switzerland, 2022a. doi: 10.1007/978-3-031-20077-9_27.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10955–10965, 2022b. doi: 10.1109/CVPR52688.2022.01069.

Yan Li, Weiwei Guo, Xue Yang, Ning Liao, Dunyun He, Jiaqi Zhou, and Wenxian Yu. Toward open vocabulary aerial object detection with CLIP-activated student-teacher learning. In *Computer Vision – ECCV 2024*, pp. 431–448. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73016-0_25.

Yuqiu Li, Jian Xue, and Ke Lu. Oriented teacher: Exploiting spatial resolution disparity to enhance semi-supervised aerial object detection. *IEEE Open Journal of Signal Processing*, 7:472–481, 2026. doi: 10.1109/OJSP.2026.3685499.

Dingkang Liang, Wei Hua, Chunsheng Shi, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. SOOD++: Leveraging unlabeled data to boost oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):840–858, 2026. doi: 10.1109/TPAMI.2025.3611519. URL https://doi.org/10.1109/TPAMI.2025.3611519.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 936–944, 2017. doi: 10.1109/CVPR.2017.106.

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*, pp. 15579–15588, 2023. doi: 10.1109/CVPR52729.2023.01495.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3390838.

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. URL https://openreview.net/forum?id=MJIve1zgR_.

Seyed Mojtaba Marvasti-Zadeh, Nilanjan Ray, and Nadir Erbilgin. Training-based model refinement and representation disagreement for semi-supervised object detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2245–2254, 2024. doi: 10.1109/WACV57701.2024.00225.

Lingqiang Meng, Shengkai Gao, Xin Lyu, Hongmin Gao, Zhizhong Liu, Yiwei Fang, Zhennan Xu, and Xin Li. Jointly enhancing representation and supervision for semisupervised rotated object detection of remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 19:25850–25870, 2026. doi: 10.1109/JSTARS.2026.3718516.

Jiancheng Pan, Yanxing Liu, Yuqian Fu, Muyuan Ma, Jiahao Li, Danda Pani Paudel, Luc Van Gool, and Xiaomeng Huang. Locate anything on earth: Advancing open-vocabulary object detection for remote sensing community. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (6):6281–6289, 2025. doi: 10.1609/aaai.v39i6.32672.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

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*, pp. 9626–9635, 2019. doi: 10.1109/ICCV.2019.00972.

Kai Wang, Zhifeng Xiao, Qiao Wan, Fanfan Xia, Pin Chen, and Deren Li. Global focal learning for semi-supervised oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024a. doi: 10.1109/TGRS.2024.3438844.

Wenyong Wang, Yuanzheng Cai, and Tao Wang. SRODET: Semi-supervised remote sensing object detection with dynamic pseudo-labeling. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025. doi: 10.1109/LGRS.2025.3544807.

Yuhao Wang, Lifan Yao, Xinye Zhang, Jiayun Song, and Haopeng Zhang. Mining oriented information for semi-supervised object detection in remote sensing images. In *2024 International Joint Conference on Neural Networks*, pp. 1–8, 2024b. doi: 10.1109/IJCNN60899.2024.10650864.

Wenhao Wu, Hau-San Wong, and Si Wu. Pseudo-siamese teacher for semi-supervised oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3380645.

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*, pp. 3974–3983, 2018. doi: 10.1109/CVPR.2018.00418.

Tao Xia, Wei Jing, and Qi Wang. Adjacent teacher: Semi-supervised oriented object detection leveraging adjacent spatial consistency prior in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025. doi: 10.1109/TGRS.2025.3579430.

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*, pp. 3040–3049, 2021. doi: 10.1109/ICCV48922.2021.00305.

Zewei Yang, Jie Feng, Junpeng Zhang, Weisheng Dong, and Ronghua Shang. Dual-criterion pseudo label selection for semi-supervised oriented object detection on remote sensing images. In *2025 IEEE International Geoscience and Remote Sensing Symposium*, pp. 8506–8510, 2025. doi: 10.1109/IGARSS55030.2025.11243394.

Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. DetCLIP: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9125–9138, 2022. doi: 10.52202/068431-0663.

Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. DetCLIPv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23497–23506, 2023. doi: 10.1109/CVPR52729.2023.02250.

Lei Zhang, Yuxuan Sun, and Wei Wei. Mind the gap: Polishing pseudo labels for accurate semi-supervised object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37 (3):3463–3471, 2023. doi: 10.1609/aaai.v37i3.25455.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024. doi: 10.1109/TGRS.2024.3449154.

Tong Zhao, Qiang Fang, Shuohao Shi, and Xin Xu. Density-guided dense pseudo label selection for semi-supervised oriented object detection. In *2024 IEEE International Conference on Image Processing*, pp. 1092–1098, 2024. doi: 10.1109/ICIP51287.2024.10647322.

Tong Zhao, Qiang Fang, and Xin Xu. Denser teacher: Rethinking dense pseudo-label for semi-supervised oriented object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4549–4559, 2025a. doi: 10.1109/TCSVT.2024.3518452.

Ye Zhao, Yong Zhu, Yuxi Gong, Xueliang Liu, and Liangfeng Xu. CBLC-SOOD: Contrastive background and label correction for semi-supervised oriented object detection. *Multimedia Systems*, 31(6):438, 2025b. doi: 10.1007/s00530-025-02028-5.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16772–16782, 2022. doi: 10.1109/CVPR52688.2022.01629.

Beichen Zhou, Qi Bi, Jian Ding, and Gui-Song Xia. Boosting fine-grained oriented object detection via text features. In *Pattern Recognition*, pp. 109–125. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-78444-6_8.

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 *Computer Vision – ECCV 2022*, pp. 35–50. Springer Nature Switzerland, 2022a. doi: 10.1007/978-3-031-20077-9_3.

Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, Wenwei Zhang, and Kai Chen. MMRotate: A rotated object detection benchmark using PyTorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 7331–7334, 2022b. doi: 10.1145/3503161.3548541.

Ye Zhu, Yan Bai, Shixin Cen, and Cuihong Xue. Towards semi-supervised adaptive oriented object detection. *Remote Sensing Letters*, 16(6):665–675, 2025. doi: 10.1080/2150704X.2025.2492903.