Towards Sparsely Annotated Open-World Object Detection

HeeJu Han, AJeong Kim, and Jinsun Park*

Pusan National University, Republic of Korea
{hhjjoa2817, jeonk636, jspark}@pusan.ac.kr

Abstract. Real-world object detection operates under ambiguous supervision, where unlabeled regions may correspond to missing annotations of known objects or genuinely unknown categories. These challenges have been addressed separately in Sparsely Annotated Object Detection (SAOD) and Open-World Object Detection (OWOD). In practice, their co-occurrence remains an open problem. To address this problem, we introduce Sparsely Annotated Open-World Object Detection (SA-OWOD), a new task that jointly considers sparse supervision and the presence of unseen categories. We propose Dual-Perspective Object Discovery (DPOD), a unified framework that jointly models unlabeled known and unknown instances via two complementary mechanisms. The Known Target Recovery Module (KTRM) recovers supervision for unlabeled known instances and explicitly regularizes the feature space to separate known and unknown representations. Complementarily, the Dual-Disagreement Target Generator (DDTG) identifies reliable unknown candidates through cross-view semantic inconsistency. By integrating these modules, DPOD resolves contradictory supervision signals caused by ambiguous unlabeled regions. As a result, it prevents misclassification between known and unknown objects and stabilizes the decision boundaries. Experimental results on sparsely annotated open-world benchmarks demonstrate that the proposed method outperforms existing open-world detection methods, particularly in detecting unknown objects. The code is publicly available at: https://github.com/HelloHeeju/SA-OWOD

Keywords: Open-World Object Detection · Sparsely Annotated Object Detection · Ambiguous Supervision

1 Introduction

Object detection inevitably operates under incomplete and semantically open conditions. In practice, many object instances remain unlabeled due to annotation costs and human limitations. However, most existing object detection methods [1, 2, 8, 13, 17, 29] rely on two strong assumptions: (1) training datasets provide fully annotated labels without missing objects, and (2) the category set is closed, with no novel classes appearing after training. While recent progress

* Corresponding author

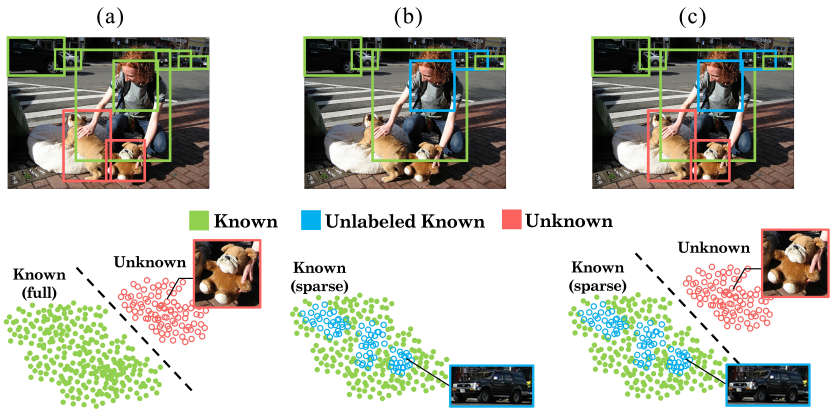

Fig. 1: Object detection paradigms for discovering unlabeled data. (a) *Open-World Object Detection* aims to identify novel objects as unknown, assuming full annotation data. (b) *Sparsely Annotated Object Detection* focuses on recovering unlabeled known objects under sparse annotation settings, while assuming a closed set of categories. (c) *Sparsely Annotated Open-World Object Detection* considers sparse annotations and open-world scenarios simultaneously, where unlabeled regions may correspond to either unlabeled known objects or unknown objects.

has been driven by large-scale datasets [3, 7, 15, 40, 43] and advances in deep neural architectures, these assumptions rarely hold in realistic scenarios.

Open-World Object Detection (OWOD) [5,14,24,26,49] aims to not only detect known objects but also identify unseen ones as *unknown*. To this end, existing OWOD methods adopt diverse strategies to model unknown objects, such as learning class-agnostic objectness scores and exploiting energy-based uncertainty measures. Nevertheless, these methods still depend on densely annotated training data, as illustrated in Fig. 1 (a). However, with sparse annotations where unlabeled known objects may exist, they are treated as background during training. As a result, the detector may incorrectly suppress valid known instances or confuse them with unknown objects, which blurs the decision boundary between known and unknown classes and degrades the accuracy of both.

On the other hand, Sparsely Annotated Object Detection (SAOD) [16,22,33, 38, 41] addresses object detection under incomplete annotation settings, where only a subset of object instances is labeled per image during training (*i.e.*, sparse annotations). In this setting, missing annotations cause detectors to incorrectly treat unlabeled objects as background regions, leading to false-negative supervision during training. To address this issue, several studies [34,36,47] attempt to compensate for annotation sparsity by generating pseudo-labels from unlabeled regions. However, these methods are fundamentally developed under a closed-world assumption, as shown in Fig. 1 (b). Under this assumption, every object belongs to a predefined set of known categories. As a result, semantically novel

objects outside the target classes are treated as background and therefore cannot be discovered.

Although SAOD and OWOD have been studied independently, real-world data contains both unlabeled known and unknown objects simultaneously. From the detector's perspective, these instances appear as indistinguishable unlabeled regions during training. Without explicit supervision, the model cannot distinguish between unlabeled instances of known classes and unseen objects, leading to ambiguous training signals. To address this problem, we introduce a new task, *Sparsely Annotated Open-World Object Detection (SA-OWOD)*, which assumes that unlabeled regions may contain both missing annotations of known categories and objects from unseen classes. As illustrated in Fig. 1 (c), SA-OWOD captures scenarios in which both types of unlabeled instances co-exist within a single image. Unlike OWOD and SAOD, which address these challenges in isolation, SA-OWOD explicitly models their interplay.

Solving SA-OWOD requires a framework that can both recover unlabeled known objects and identify unknown objects. For this purpose, we introduce *Dual-Perspective Object Discovery (DPOD)*. DPOD comprises two complementary components: the *Known Target Recovery Module (KTRM)* and the *Dual-Disagreement Target Generator (DDTG)*. KTRM recovers unlabeled known objects through pseudo-labeling and regularizes the decision boundary between known and unknown categories through feature-level alignment. Complementing this, DDTG identifies additional unknown candidates by exploiting cross-view semantic disagreement. Proposals with high objectness but low logit similarity are treated as reliable unknowns. By coupling pseudo-label recovery with disagreement-driven unknown discovery and feature-level regularization, DPOD resolves the inherent ambiguity of unlabeled regions in SA-OWOD.

The main contributions of our work are summarized as follows:

- To the best of our knowledge, we are the first to formulate the Sparsely Annotated Open-World Object Detection (SA-OWOD) task. To address this challenge, we introduce a novel framework, DPOD.
- We propose two complementary modules, KTRM and DDTG, to resolve the ambiguity of unlabeled regions in SA-OWOD. KTRM recovers missing known objects to stabilize pseudo-label learning and refine decision boundaries, while DDTG discovers additional unknown instances via cross-view semantic disagreement.
- We construct sparsely annotated open-world benchmarks and an evaluation protocol that explicitly captures the dual nature of unlabeled regions. They allow quantitative evaluation of both missing known object recovery and unknown object discovery.

2 Related Work

Object detection has achieved remarkable progress, evolving from two-stage frameworks [11,30] and efficient one-stage detectors [19,29] to recent transformer-based architectures [2,48]. Despite their strong performance, these methods oper-

ate under a closed-world assumption — they recognize only predefined categories and treat all unlabeled or unknown objects as background. Consequently, these methods rely on static, fully annotated datasets. This limits their applicability to real-world scenarios, where annotations are often incomplete and novel categories continually emerge.

2.1 Open-World Object Detection (OWOD)

Open-World Object Detection (OWOD) [23, 37, 39, 44, 46] aims to not only recognize known object categories but also identify novel (*i.e.*, unknown) objects and incrementally learn them over time.

Pseudo-labeling methods identify unknown objects and generate pseudo-labels as additional supervision. ORE [14] introduces the OWOD setting and generates unknown pseudo-labels from high-objectness proposals. It further enhances unknown modeling using contrastive clustering and energy-based identification. OW-DETR [9] extends Deformable DETR [48] with attention-guided pseudo-labeling and explicit supervision to distinguish unknown objects from background regions.

Class-agnostic methods focus on objectness rather than class-specific supervision, treating both known and unknown objects as foreground. PROB [49] introduces a probabilistic objectness head to generalize from known to unseen categories. RandBox [35] mitigates the bias of label-dependent proposal generation by adopting randomly generated region proposals. OrthogonalDet [31] decouples objectness and classification by enforcing orthogonality between objectness and class features. CROWD [25] discovers diverse unknown instances via a submodular gain objective and then learns disentangled representations of known and unknown classes to reduce confusion.

However, despite their effectiveness, most existing OWOD methods still depend on densely annotated data. Under sparse annotation settings, unlabeled known objects obscure the boundary between known and unknown categories, leading to degraded detection performance.

2.2 Sparsely Annotated Object Detection (SAOD)

Sparsely Annotated Object Detection (SAOD) [10, 28, 32, 42, 45] addresses object detection when only a subset of object instances is annotated per image, resulting in sparse supervision. In such scenarios, detectors incorrectly interpret unlabeled objects as background regions, introducing false-negative supervision during training.

To alleviate this issue, several studies leverage pseudo-labeling to compensate for missing annotations. uDenseTeacher [47] utilizes dense pseudo-labels generated from the teacher's raw dense predictions instead of sparse box-level pseudo-labels with post-processing. Co-mining [34] jointly trains two detectors that exchange each other's high-confidence predictions. Co-Student [36] adopts a collaborative strong–weak student framework. The two students leverage each

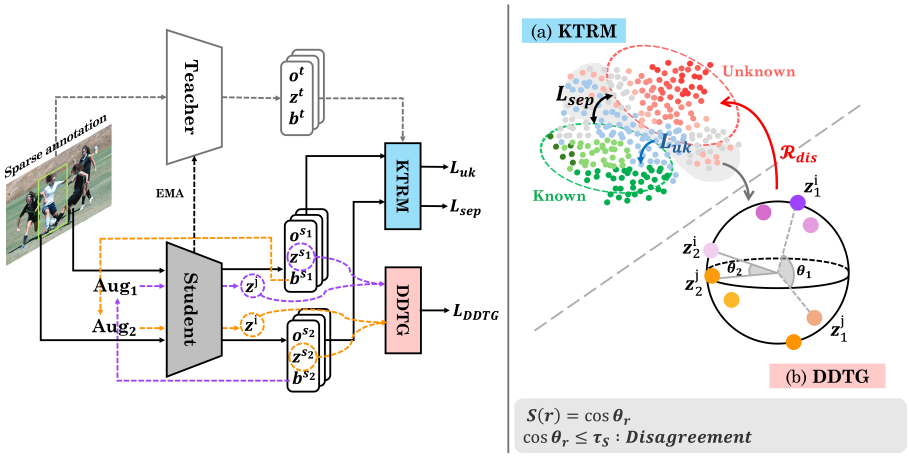

Fig. 2: Overview of DPOD. (a) KTRM recovers supervision for unlabeled known objects and enforces known–unknown feature separation. (b) DDTG identifies additional unknown candidates via cross-view semantic inconsistency. The two signals jointly strengthen known–unknown separation and stabilize feature representations.

other's predictions, while a teacher model refines them to provide more reliable supervision.

However, these methods are designed under a closed-world assumption, where every object belongs to a predefined class set. As a result, objects from previously unseen categories are implicitly treated as background, preventing their detection.

3 Method

3.1 SA-OWOD Problem Definition

In this work, we introduce a new task, SA-OWOD, which extends OWOD to a sparsely annotated scenario, as shown in Fig. 2. In the standard OWOD formulation [14], a detector is trained to recognize known classes, identify unseen objects as *unknown*, and incrementally incorporate them over time. Unlike conventional OWOD, SA-OWOD assumes partial supervision within known classes: not all instances of known categories are annotated during training. Consequently, unlabeled known objects and truly unknown instances coexist without labels, yet must be assigned distinct semantic roles. This ambiguity constitutes the core challenge of SA-OWOD.

Formally, the SA-OWOD is defined as follows: At stage t, a detector f_t is trained on a dataset $\mathcal{D}_t = \{\mathcal{I}_t, \mathcal{L}_t\}$, where $\mathcal{I}_t$ and $\mathcal{L}_t$ denote a set of training images and their corresponding partial annotations, respectively. The annotations are provided only for the known class set $\mathcal{K}_t = \{1, 2, \ldots, C\}$. However,

not all instances belonging to $\mathcal{K}_t$ are annotated; a subset of known-class objects may remain unlabeled in $\mathcal{I}_t$. In addition, objects from the unknown class set $\mathcal{U} = \{C+1, C+2, \dots\}$ may appear without labels during training. Consequently, unlabeled regions may correspond to either missing annotations of known classes or objects from unknown categories.

Similar to OWOD, the problem follows an incremental learning protocol. At each stage, a subset of previously detected unknown classes $\bar{\mathcal{U}}_t \subset \mathcal{U}$ is annotated and incorporated into the known class set $\mathcal{K}_{t+1} = \mathcal{K}_t \cup \bar{\mathcal{U}}_t$. Due to memory and computational constraints, the model is updated from f_t to f_{t+1} using limited samples from previously known classes rather than retraining from scratch. This process is repeated over T stages, requiring the detector to continuously discover novel objects while mitigating catastrophic forgetting.

In summary, SA-OWOD requires the detector to simultaneously: (1) recover supervision for unlabeled known instances and (2) detect truly unknown objects.

3.2 Known Target Recovery Module (KTRM)

In SA-OWOD, unlabeled regions may correspond to known objects that were missed during annotation. If left unaddressed, these unlabeled known instances are treated as background during training. This introduces false-negative supervision and blurs the decision boundary between known and unknown categories. KTRM is designed to explicitly resolve this ambiguity through two complementary mechanisms: pseudo-label recovery for unlabeled known objects and representation-level separation between unlabeled known and unknown proposals. In sparsely annotated open-world settings, supervision for known classes is inherently incomplete, which makes classifier-based pseudo-labeling unreliable due to ambiguous decision boundaries. To alleviate this issue, we adopt a class-agnostic open-world detection framework based on CROWD [25], which produces object-centric proposals.

Based on these object-centric proposals, we recover supervision for unlabeled known objects using a pseudo-label generation strategy inspired by Co-Student [36]. A student detector first generates predictions from multiple augmented views of the same image. These predictions are then refined by a teacher model through consistency-aware filtering, producing a reliable unlabeled known proposal set $\mathcal{R}_{uk}$. We treat the elements of $\mathcal{R}_{uk}$ as pseudo-labels and combine them with the sparse ground-truth annotations to compensate for missing labels in known categories. Each pseudo-label $p \in \mathcal{R}_{uk}$ provides a class label y_p^{uk} and a bounding box $\mathbf{b}_p^{uk}$. After that, each pseudo-label p searches for the object proposal r_p with the largest IoU and they are used for the training via classification loss ℓ_{cls} and regression loss ℓ_{reg}. The pseudo-label supervision is defined as:

$$\mathcal{L}_{uk} = \frac{1}{|\mathcal{R}_{uk}|} \sum_{p \in \mathcal{R}_{uk}} \Big(\ell_{cls}\big(\mathbf{z}_{r_p}, y_p^{uk}\big) + \ell_{reg}\big(\mathbf{b}_{r_p}, \mathbf{b}_p^{uk}\big) \Big), \tag{1}$$

where $\mathbf{z}_{r_p}$ and $\mathbf{b}_{r_p}$ denote class logits and bounding box of r_p, respectively.

While pseudo-label recovery compensates for missing supervision, feature representations of unlabeled known objects can still overlap with unknown instances in the embedding space. To enforce representation-level separation, we adopt a conditional gain objective inspired by the Facility-Location formulation in CROWD. We adopt the Facility-Location formulation because it encourages each proposal to be separated from its nearest unknown counterpart rather than considering all unknown proposals simultaneously. In contrast, distance-based objectives aggregate similarities across multiple unknown instances, which tends to blur local decision boundaries and weaken fine-grained discrimination.

Let $\mathcal{R}_k$, $\mathcal{R}_{uk}$, and $\mathcal{R}_u$ denote labeled known (from sparse ground-truth annotations), unlabeled known (from pseudo-labeling), and unknown proposal sets, respectively. We define the unified known set as:

$$\mathcal{A} = \mathcal{R}_k \cup \mathcal{R}_{uk}. \tag{2}$$

Let $\mathcal{V}$ denote the set of all proposals. The separation objective is formulated as a facility-location conditional gain:

$$\mathcal{L}_{sep} = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \max\Big(0, \max_{j \in \mathcal{A}} s_{ij} - \eta \max_{j \in \mathcal{R}_u} s_{ij}\Big), \tag{3}$$

where s_{ij} denotes a similarity metric between i-th and j-th proposals, and η is a weighting factor that encourages separation between known and unknown features.

Minimizing $\mathcal{L}_{sep}$ encourages known proposals to move away from unknown regions in the embedding space, thereby reducing feature entanglement and stabilizing decision boundaries under sparse supervision.

The overall KTRM objective is defined as:

$$\mathcal{L}_{KTRM} = \alpha_{uk}\, \mathcal{L}_{uk} + \alpha_{sep}\, \mathcal{L}_{sep}, \tag{4}$$

where α_{uk} and α_{sep} are balancing hyperparameters.

3.3 Dual-Disagreement Target Generator (DDTG)

In open-world settings, unknown objects often manifest as regions where semantic predictions are inherently unstable [31]. Unlike known categories, unknown objects tend to reside near ambiguous decision boundaries. Motivated by this observation, we introduce the Dual-Disagreement Target Generator (DDTG), summarized in Algorithm 1. DDTG identifies unknown object candidates by measuring cross-view semantic inconsistency. Such inconsistency typically arises from unstable regions in the classification space associated with unknown objects.

Given two views of the same image, the detector produces class logits for each object proposal. We compare the class logit vectors obtained from the two views to identify prediction disagreement. However, reliable comparison requires that the corresponding regions of interest (RoIs) refer to the same spatial region.

Algorithm 1 Procedure for Selecting Unknown Candidates in DDTG

Require: Detector f_θ, two views $\{x^{s_1}, x^{s_2}\}$, objectness threshold τ_o, similarity threshold τ_s
Ensure: Disagreement proposal set $\mathcal{R}_{dis}$

```
Initialize R_dis ← ∅
for each ordered pair (x^i, x^j) ∈ {(x^{s1}, x^{s2}), (x^{s2}, x^{s1})} do
    Obtain region proposals R^a from x^a
    for each proposal r ∈ R^a do
        Project r to x^j
        Extract logits z_r^i = f_θ(x^i, r), z_r^j = f_θ(x^j, r)
        Compute similarity S(r) = (z_r^i)^⊤ z_r^j / (‖z_r^i‖_2 ‖z_r^j‖_2)
        if o_r ≥ τ_o and S(r) ≤ τ_s then
            R_dis ← R_dis ∪ {r}
        end if
    end for
end for
return R_dis
```

Naively matching proposals by intersection-over-union (IoU) can be unstable, as even minor localization shifts may lead to substantially different RoI features. To ensure consistent cross-view comparison, we generate each RoI from a reference view and project it to the other view to extract geometrically aligned RoI features. Prediction disagreement is then quantified by computing the cosine similarity between the aligned class logit vectors.

Given a proposal r, let $\mathbf{z}_r^i \in \mathbb{R}^C$ and $\mathbf{z}_r^j \in \mathbb{R}^C$ denote the classification logits predicted from two different views. We measure their semantic consistency using cosine similarity:

$$\mathcal{S}(r) = \frac{\left(\mathbf{z}_r^i\right)^\top \mathbf{z}_r^j}{\|\mathbf{z}_r^i\|_2 \, \|\mathbf{z}_r^j\|_2}. \tag{5}$$

The set of disagreement proposals is defined as:

$$\mathcal{R}_{dis} = \Big\{ r \in \mathcal{R} \;\Big|\; o_r \geq \tau_o,\ \mathcal{S}(r) \leq \tau_s \Big\}, \tag{6}$$

where o_r denotes the objectness score of proposal r, τ_o is the objectness threshold, and τ_s is the similarity threshold. Only proposals that are sufficiently object-like but exhibit low semantic consistency are treated as disagreement candidates. Similar to KTRM, we apply both classification and separation objectives to the disagreement set in order to enforce consistent unknown prediction. The resulting DDTG loss is defined as:

$$\begin{aligned} \mathcal{L}_{DDTG} = \beta_{cls} \frac{1}{|\mathcal{R}_{dis}|} \sum_{d \in \mathcal{R}_{dis}} \ell_{cls}\big(\mathbf{z}_{r_d}, y_d^{dis}\big) \\ - \beta_{sep} \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \max\Big(0, \max_{j \in \mathcal{A}} s_{ij} - \eta \max_{j \in \mathcal{R}_{dis}} s_{ij}\Big), \end{aligned} \tag{7}$$

Table 1: Task composition in open-world evaluation protocol. The semantics of each task and the number of images and instances (objects) across splits are shown.

	Task 1	Task 2	Task 3	Task 4
Semantic split	VOC Classes	Outdoor, Accessories, Appliance, Truck	Sports, Food	Electronic, Indoor, Kitchen, Furniture
# training images	16,551	45,520	39,402	40,260
# test images	4,952	1,914	1,642	1,738
# train instances	47,223	113,741	114,452	138,996
# test instances	14,976	4,966	4,826	6,039

where y_d^{dis} denotes the unknown class label, and $\mathbf{z}_{r_d}$ denotes the class logits of the proposal r_d that best matches the disagreement candidate d. β_{cls} and β_{sep} are balancing hyperparameters.

DDTG provides an additional supervisory signal for unknown object learning by modeling cross-view semantic instability. While objectness filtering ensures that candidate regions correspond to potential objects, it does not capture ambiguity in class predictions under view perturbations. By explicitly measuring disagreement between aligned logits, DDTG complements objectness-based supervision and enhances the robustness of unknown object discovery.

Accordingly, the overall loss for unlabeled object supervision is defined as the combination of the KTRM loss and the DDTG loss:

$$\mathcal{L}_{DPOD} = \gamma_{KTRM}\,\mathcal{L}_{KTRM} + \gamma_{DDTG}\,\mathcal{L}_{DDTG}, \tag{8}$$

where γ_{KTRM} and γ_{DDTG} are balancing hyperparameters.

4 Experiments

4.1 Datasets and Sparse Annotation Settings

We utilize the Open-World Object Detection dataset [14], which is constructed from Pascal VOC [6] and MS-COCO [20]. Following the standard OWOD dataset protocol, we organize all VOC classes as the first task T_1, while the remaining 60 COCO classes are divided into three successive tasks. This division introduces semantic drifts between tasks (see Tab. 1). During the training of task T_t, classes from previous tasks $\{T_\tau : \tau < t\}$ are treated as *known*, while classes from future tasks $\{T_\tau : \tau > t\}$ are treated as *unknown*. This setup enables progressive learning under an open-world scenario, where the model incrementally encounters new classes while retaining knowledge of previously learned ones.

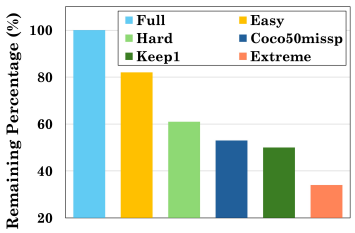

Fig. 3: Remaining annotation ratio under each sparse configuration (all tasks aggregated).

To simulate sparse annotation scenarios, we consider five training configurations [18, 45] designed to reduce available annotations. Unlike conventional sparse annotation settings that remove annotations globally across the entire dataset, our framework follows an open-world protocol in which different tasks introduce different sets of classes. Accordingly, annotation removal is applied independently within each task. For example, under the Easy setting, one annotation is removed per image for each task individually: one annotation is removed for classes in Task 1, another for classes in Task 2, and so on. This ensures that each task observes its own sparse annotation setting, while images shared across tasks may have multiple annotations removed corresponding to the task-specific classes.

Specifically, five configurations are defined based on the number of retained annotations per image per task:

- **Easy**: Remove a single annotation, retaining at least one annotation per class in the current task; approximately 18.3% of all labels are removed.
- **Hard**: Remove half of the annotations, retaining at least one annotation per class in the current task; approximately 38.8% of all labels are removed.
- **Coco50missp**: Remove half of the annotations; approximately 46.7% of labels are removed. Unlike the Hard setting, this does not guarantee at least one annotation per class in each image.
- **Keep1**: Keep only one annotation for each category present in an image; approximately 50.0% of labels are removed.
- **Extreme**: Retain only a single annotation; approximately 65.8% of labels are removed.

For evaluation, we adopt the Pascal VOC test split and the MS COCO validation split.

4.2 Implementation Details

In this study, we adopt a Faster R-CNN [30] as the baseline [25]. We employ a ResNet50 [12] backbone pre-trained on ImageNet [4]. The network is optimized using AdamW [21] with a learning rate of 2.5×10^{-5} and a weight decay of 1×10^{-4}. For each task, the model is trained for 15,000 iterations, followed by 15,000 iterations of incremental learning, resulting in a total of 30,000 iterations (approximately 9 epochs). All experiments are conducted on four NVIDIA RTX A6000 GPUs with a batch size of 12. All weighting coefficients α_{uk}, α_{sep}, β_{cls}, β_{sep}, γ_{KTRM}, and γ_{DDTG} are set to 1, and the facility-location parameter η is set to 1. During inference, predictions are obtained from 10,000 pre-defined bounding boxes, after which Non-Maximum Suppression (NMS) [27] is applied to remove redundant detections. Our model comprises 106.2M parameters with a computational cost of 965.4 GFLOPs, running at 2.52 FPS with a latency of 396.1 ms.

Table 2: Sparsely Annotated Open-World Object Detection (SA-OWOD) results. Existing open-world detectors are benchmarked on sparse annotation datasets for fair comparison. K-mAP and U-Recall denote Known mAP and Unknown Recall, respectively.

Set	Method	Task 1		Task 2		Task 3		Task 4
		K-mAP	U-Recall	K-mAP	U-Recall	K-mAP	U-Recall	K-mAP
Full	RandBox [35]	61.8	10.6	45.3	6.3	39.4	7.8	35.4
	PROB [49]	59.5	19.4	44.0	17.4	36.0	19.6	31.5
	OrthogonalDet [31]	61.3	24.6	47.0	26.3	41.3	29.1	37.9
	CROWD [25]	61.7	57.9	47.8	53.6	42.5	69.6	38.5
Easy	RandBox	49.91	4.90	38.36	4.30	33.12	4.60	29.79
	PROB	52.83	18.27	38.35	15.35	32.64	19.1	29.06
	OrthogonalDet	56.19	17.05	**41.91**	22.51	36.73	25.13	**34.60**
	CROWD	53.28	45.53	38.24	30.22	35.43	53.34	32.28
	Ours	**56.48**	**55.34**	39.53	**52.47**	**36.90**	**63.68**	32.15
Hard	RandBox	49.96	2.69	27.66	3.62	29.86	3.72	27.26
	PROB	50.18	18.26	36.24	13.91	30.98	17.57	27.01
	OrthogonalDet	53.14	19.19	**38.03**	18.28	32.87	24.44	**32.14**
	CROWD	49.29	49.80	34.83	33.35	33.63	50.80	29.91
	Ours	**54.09**	**51.85**	33.36	**48.25**	**35.79**	**64.00**	31.15
Coco50missp	RandBox	45.43	0.63	28.02	0.76	16.93	1.47	21.09
	PROB	48.84	18.27	34.58	14.46	28.96	16.96	25.65
	OrthogonalDet	**54.38**	15.41	**35.56**	25.16	**31.26**	29.24	**28.56**
	CROWD	49.90	43.91	32.03	24.14	29.92	41.12	26.13
	Ours	53.78	**51.39**	30.73	**47.69**	30.90	**59.00**	24.74
Keep1	RandBox	44.35	2.95	31.34	3.70	28.72	4.18	25.05
	PROB	48.81	17.84	34.67	14.06	29.59	16.95	26.47
	OrthogonalDet	**50.47**	14.07	**35.00**	16.60	31.58	19.71	**30.28**
	CROWD	46.45	44.59	32.76	33.27	32.16	52.07	28.34
	Ours	47.97	**48.03**	33.09	**48.28**	**32.37**	**52.22**	28.21
Extreme	RandBox	36.62	0.68	21.93	1.50	19.55	2.39	17.95
	PROB	37.78	17.86	26.09	14.83	23.41	17.60	19.38
	OrthogonalDet	41.39	12.41	**26.20**	14.77	23.30	24.62	**22.31**
	CROWD	34.37	40.66	24.84	31.17	24.46	48.64	20.47
	Ours	**45.37**	**45.56**	25.98	**44.32**	**26.58**	**57.53**	20.19

4.3 Main Results

SA-OWOD Performance. Table 2 compares existing OWOD methods under sparse annotations. All methods are retrained under identical sparse annotation protocols for fair comparison. K-mAP denotes the mean Average Precision over known categories, and U-Recall evaluates unknown object discovery. As annotation sparsity increases, prior OWOD approaches exhibit substantial degradation, particularly in U-Recall. In contrast, our method achieves superior U-Recall while maintaining competitive K-mAP across all sparsity levels.

In the Easy setting, our method improves U-Recall by 9.81%p over CROWD and by 38.29%p over OrthogonalDet in Task 1, achieving the best K-mAP in Tasks 1 and 3. These results demonstrate that our framework improves unknown discovery without sacrificing known-class detection. As illustrated in Fig. 4,

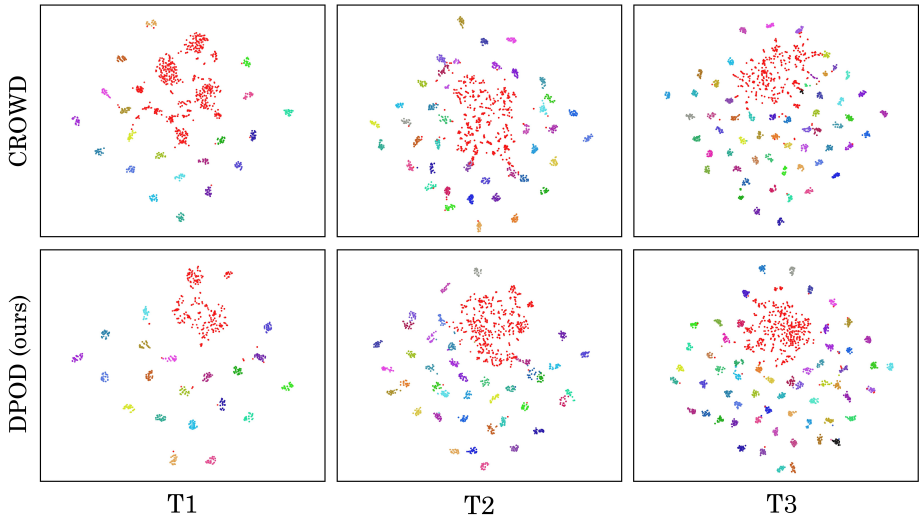

Fig. 4: t-SNE visualization of feature embeddings. The distributions of known and unknown samples are shown for Tasks 1–3. Known classes are represented by different colors, while unknown objects are shown in red. Our method shows clearer separation between known and unknown samples compared to the baseline.

the baseline tends to mix unknown samples with known clusters in the feature space. This phenomenon arises because sparsely annotated datasets often contain unlabeled known objects that are unintentionally treated as unknown during training, leading to ambiguous feature representations. KTRM recovers these via objectness-guided pseudo-labeling, preventing them from being incorrectly treated as unknown samples. Simultaneously, DDTG leverages cross-view semantic disagreement to identify reliable unknown candidates. By disentangling unlabeled known and unknown instances, our framework reduces the boundary confusion that often degrades prior OWOD methods under sparse supervision.

As sparsity intensifies (*i.e.*, Easy → Hard → Extreme), the performance gap widens further. In the Hard setting, our method surpasses CROWD by 14.90%p and OrthogonalDet by 29.97%p in Task 2. Under the Extreme sparsity, the unknown recall gap over OrthogonalDet exceeds 29.55%p in multiple tasks. This robustness indicates that KTRM effectively compensates for missing supervision, while DDTG prevents premature consolidation of ambiguous proposals into known categories.

Challenging Stages. In Task 2, our K-mAP is slightly lower than OrthogonalDet due to integrating novel classes under incomplete supervision, which causes unstable boundaries. However, DDTG improves robustness to ambiguity by delaying overconfident assignments under weak cross-view semantic consistency. This results in a 29.97%p gain in U-Recall over OrthogonalDet in Hard Task 2. In Extreme Task 2, OrthogonalDet achieves the highest absolute K-

mAP; however, its U-Recall drops by 43.84% compared to the fully annotated setting. In contrast, although our method exhibits a comparable K-mAP reduction, the degradation in U-Recall is limited to only 17.31%. This gap demonstrates that our framework achieves a better balance between known detection and unknown discovery under sparse annotations.

In contrast, Task 4, a later incremental stage, has few remaining unknowns. Consequently, the amount of supervision related to unknown discovery becomes limited, and the learning objective shifts toward fine-grained discrimination among many known categories. Since our framework is primarily designed to resolve ambiguity between unlabeled known and unknown objects, the relative contribution of disagreement-driven unknown modeling (DDTG) naturally decreases when unknown signals become scarce. Therefore, the smaller performance margin observed in Task 4 reflects a change in task characteristics rather than a limitation in modeling class separation. Even under this shift, our method maintains competitive performance, demonstrating stable behavior across incremental learning stages.

Discussion. The core challenge of SA-OWOD lies in resolving the ambiguity of unlabeled regions. KTRM reduces false-negative supervision caused by missing annotations, preserving K-mAP stability under sparsity. DDTG strengthens unknown modeling by explicitly regularizing semantically inconsistent proposals. The reduced U-Recall degradation and performance gap validate that explicitly modeling both mechanisms is essential for robust detection in sparsely annotated open-world environments.

4.4 Ablation Studies

We conduct ablation studies under the Hard setting to analyze the individual contributions of the proposed KTRM and DDTG, as well as the effect of key thresholds in DDTG. Moreover, we provide a qualitative comparison of the results with those of a baseline model.

Effect of KTRM and DDTG. As shown in Tab. 3, we perform an ablation study to analyze the individual contributions of the proposed KTRM and DDTG under the Hard setting. Starting from the CROWD baseline, we progressively enable each module to examine their impact on both known object detection performance (K-mAP) and unknown object recall (U-Recall).

When only KTRM is applied, improvements are mainly observed in known category detection, indicating that KTRM enhances discriminative representation learning and stabilizes training. Applying only DDTG also improves performance, suggesting that the proposed target generation mechanism contributes to boundary refinement and unknown exploration.

When both modules are jointly applied, consistent improvements are observed across both K-mAP and U-Recall. This demonstrates that KTRM and DDTG play complementary roles: KTRM improves feature quality for known categories, while DDTG facilitates unknown object exploration. The combined design achieves a better balance between known recognition and unknown discovery.

Table 3: Ablation study on KTRM and DDTG under the Hard setting.

Task 1	KTRM	DDTG	K-mAP	Δ	U-Recall	Δ
CROWD	✗	✗	49.29	–	49.80	–
CROWD+KTRM	✓	✗	51.90	+2.61	50.66	+0.86
CROWD+DDTG	✗	✓	51.08	+1.79	50.26	+0.46
Ours	✓	✓	**54.09**	**+4.80**	**51.85**	**+2.05**

Table 4: Ablation study on τ_o and τ_s in DDTG under the Hard setting.

Objectness threshold (τ_o)				Similarity threshold (τ_s)			
τ_o	τ_s	K-mAP	U-Recall	τ_o	τ_s	K-mAP	U-Recall
0.1	0.95	52.97	47.68	**0.2**	**0.95**	**54.09**	**51.85**
0.2	**0.95**	**54.09**	**51.85**	0.2	0.90	52.24	46.26
0.3	0.95	53.62	51.40	0.2	0.85	52.53	50.46

Sensitivity to Objectness and Similarity Thresholds. We further analyze the sensitivity of DDTG to the objectness threshold (τ_o) and similarity threshold (τ_s), as shown in Tab. 4. When fixing $\tau_s = 0.95$, increasing τ_o from 0.1 to 0.2 significantly improves both K-mAP and U-Recall, indicating that moderate filtering of low-objectness proposals helps suppress noise while preserving informative candidates. However, further increasing τ_o to 0.3 slightly degrades performance, suggesting that overly strict filtering may discard useful unknown candidates.

When fixing $\tau_o = 0.2$, increasing τ_s enlarges the set of proposals regarded as semantic disagreement. Although a larger disagreement set may potentially introduce noisy candidates, the results indicate that the additional proposals provide informative supervisory signals rather than degrading detection quality. We conjecture that richer disagreement cues encourage the model to learn a clearer separation between known and unknown representations, which benefits not only unknown recall but also known-class detection performance. Reducing τ_s limits the diversity of disagreement candidates, leading to consistent drops in both metrics. These findings suggest that sufficiently diverse semantic disagreement signals are important for stable and effective open-world detection.

Qualitative Results. The qualitative results in Fig. 5 visually support the qualitative improvements reported in Hard Task 3. In CROWD [25], known objects are sometimes misclassified as unknown, and some unknown instances fail to be clearly activated. In contrast, our method reduces ambiguity around object boundaries and produces more consistent detections. This suggests that our approach achieves improved stability even under incomplete supervision.

5 Conclusion

In this work, we introduced *Sparsely Annotated Open-World Object Detection (SA-OWOD)*, where unlabeled known and unknown objects co-exist. In this set-

Fig. 5: Qualitative Comparison on Hard Task 3. We compare CROWD [25] and our method in terms of known and unknown detections. Blue denotes unknown objects, and green denotes known objects.

ting, unlabeled regions are inherently ambiguous and cannot be reliably treated as either unlabeled known or unknown. To address this challenge, we proposed *Dual-Perspective Object Discovery (DPOD)*, a unified framework that explicitly separates unlabeled regions into recoverable known objects and genuinely unknown ones. By explicitly modeling and separating unlabeled known and unknown instances, DPOD effectively resolves this ambiguity. Extensive experiments demonstrate consistent improvements in both known-class detection and unknown recall. Overall, SA-OWOD provides a practical benchmark for detection under ambiguous supervision and lays the foundation for robust open-world detection in real-world scenarios.

Limitations. Despite these advances, DPOD currently relies on fixed objectness and score thresholds (τ_o, τ_s) for proposal filtering. In open-world scenarios, object characteristics vary significantly across tasks, and certain categories — such as accessories (e.g., ties) — tend to yield lower objectness scores. Enforcing a fixed threshold under such noisy conditions is suboptimal and may introduce progressive learning bottlenecks. Replacing these fixed heuristics with task-adaptive thresholds remains a promising direction for future work.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2024-00358935, 90%) and in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development grant funded by the Korea government (MSIT) (IITP-2026-RS-2023-00254177, 10%).

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6154–6162 (2018)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229. Springer (2020)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3213–3223 (2016)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 248–255. IEEE (2009)
5. Doan, T., Li, X., Behpour, S., He, W., Gou, L., Ren, L.: Hyp-ow: Exploiting hierarchical structure learning with hyperbolic distance enhances open world object detection. In: AAAI. vol. 38, pp. 1555–1563 (2024)
6. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge **88**(2), 303–338 (2010)
7. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 580–587 (2014)
9. Gupta, A., Narayan, S., Joseph, K., Khan, S., Khan, F.S., Shah, M.: Ow-detr: Open-world detection transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9235–9244 (2022)
10. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Adv. Neural Inform. Process. Syst. **31** (2018)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Int. Conf. Comput. Vis. pp. 2961–2969 (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016)
13. Howard, A.G.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
14. Joseph, K.J., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
15. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. Int. J. Comput. Vis. **128**(7), 1956–1981 (2020)
16. Lee, C., Shin, S., Park, G.M., Kim, J.U.: Multispectral pedestrian detection with sparsely annotated label. In: AAAI. vol. 39, pp. 4482–4490 (2025)
17. Lee, S., Park, J., Park, J.: Crossformer: Cross-guided attention for multi-modal object detection. Pattern Recognition Letters **179**, 144–150 (2024)
18. Li, H., Pan, X., Yan, K., Tang, F., Zheng, W.S.: Siod: Single instance annotated per category per image for object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14197–14206 (2022)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Int. Conf. Comput. Vis. pp. 2980–2988 (2017)

20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. pp. 740–755. Springer (2014)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. Int. Conf. Learn. Represent. (2019)
22. Lu, Z., Wang, C., Xu, C., Zheng, X., Cui, Z.: Progressive exploration-conformal learning for sparsely annotated object detection in aerial images. Adv. Neural Inform. Process. Syst. **37**, 40593–40614 (2024)
23. Ma, S., Wang, Y., Wei, Y., Fan, J., Li, T.H., Liu, H., Lv, F.: Cat: Localization and identification cascade detection transformer for open-world object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19681–19690 (2023)
24. Ma, Y., Li, H., Zhang, Z., Guo, J., Zhang, S., Gong, R., Liu, X.: Annealing-based label-transfer learning for open world object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11454–11463 (2023)
25. Majee, A., Gangrade, A., Iyer, R.: Looking beyond the known: Towards a data discovery guided open-world object detection. Adv. Neural Inform. Process. Syst. (2025)
26. Mullappilly, S.S., Gehlot, A.S., Anwer, R.M., Khan, F.S., Cholakkal, H.: Semi-supervised open-world object detection. In: AAAI. vol. 38, pp. 4305–4314 (2024)
27. Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: Int. Conf. Pattern Recog. vol. 3, pp. 850–855. IEEE (2006)
28. Niitani, Y., Akiba, T., Kerola, T., Ogawa, T., Sano, S., Suzuki, S.: Sampling techniques for large-scale object detection from sparsely annotated objects. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6510–6518 (2019)
29. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 779–788 (2016)
30. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inform. Process. Syst. **28** (2015)
31. Sun, Z., Li, J., Mu, Y.: Exploring orthogonality in open world object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17302–17312 (2024)
32. Suri, S., Rambhatla, S., Chellappa, R., Shrivastava, A.: Sparsedet: Improving sparsely annotated object detection with pseudo-positive mining. In: Int. Conf. Comput. Vis. pp. 6770–6781 (2023)
33. Wang, H., Liu, L., Zhang, B., Zhang, J., Zhang, W., Gan, Z., Wang, Y., Wang, C., Wang, H.: Calibrated teacher for sparsely annotated object detection. In: AAAI. vol. 37, pp. 2519–2527 (2023)
34. Wang, T., Yang, T., Cao, J., Zhang, X.: Co-mining: Self-supervised learning for sparsely annotated object detection. In: AAAI. vol. 35, pp. 2800–2808 (2021)
35. Wang, Y., Yue, Z., Hua, X.S., Zhang, H.: Random boxes are open-world object detectors. In: Int. Conf. Comput. Vis. pp. 6233–6243 (2023)
36. Wu, L., Han, J., Zheng, Z., Wang, X.: Co-student: Collaborating strong and weak students for sparsely annotated object detection. In: Eur. Conf. Comput. Vis. pp. 459–475. Springer (2024)
37. Wu, Y., Zhao, X., Ma, Y., Wang, D., Liu, X.: Two-branch objectness-centric open world detection. In: Proc. Int. Workshop Hum.-Centric Multimedia Anal. pp. 35–40 (2022)
38. Wu, Z., Bodla, N., Singh, B., Najibi, M., Chellappa, R., Davis, L.S.: Soft sampling for robust object detection (2019)

39. Wu, Z., Lu, Y., Chen, X., Wu, Z., Kang, L., Yu, J.: Uc-owod: Unknown-classified open world object detection. In: Eur. Conf. Comput. Vis. pp. 193–210. Springer (2022)
40. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3974–3983 (2018)
41. Yao, S., Liu, Y., Jia, Q., Chen, S., Zhuo, W.: As pseudo-label free as possible: Leveraging adaptive feature generation for sparsely annotated object detection. In: AAAI. vol. 39, pp. 9418–9426 (2025)
42. Yoon, J., Hong, S., Choi, M.K.: Semi-supervised object detection with sparsely annotated dataset. In: IEEE Int. Conf. Image Process. pp. 719–723. IEEE (2021)
43. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2636–2645 (2020)
44. Yu, J., Ma, L., Li, Z., Peng, Y., Xie, S.: Open-world object detection via discriminative class prototype learning. arXiv preprint arXiv:2302.11757 (2023)
45. Zhang, H., Chen, F., Shen, Z., Hao, Q., Zhu, C., Savvides, M.: Solving missing-annotation object detection with background recalibration loss. In: ICASSP. pp. 1888–1892. IEEE (2020)
46. Zhang, S., Ni, Y., Du, J., Xue, Y., Torr, P., Koniusz, P., van den Hengel, A.: Open-world objectness modeling unifies novel object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 30332–30342 (2025)
47. Zhou, H., Ge, Z., Liu, S., Mao, W., Li, Z., Yu, H., Sun, J.: Dense teacher: Dense pseudo-labels for semi-supervised object detection. In: Eur. Conf. Comput. Vis. pp. 35–50. Springer (2022)
48. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection (2021)
49. Zohar, O., Wang, K.C., Yeung, S.: Prob: Probabilistic objectness for open world object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11444–11453 (2023)