

Calibrate Before Adapt: Training-Free Pseudo-Label Calibration for Semi-Supervised Cross-Domain Few-Shot Object Detection

Xin Yang^{1*}, Fei Zhou^{2*}, Lei Zhang³, and Wei Wei^{3✉}

¹ National Elite Institute of Engineering,
Northwestern Polytechnical University, China
yxx_@mail.nwpu.edu.cn

² MMLab, The Chinese University of Hong Kong, China
zhoufeiyinge@gmail.com

³ School of Computer Science, Northwestern Polytechnical University, China
nwpuzhanglei@nwpu.edu.cn, weiweinwpu@nwpu.edu.cn

Abstract. Semi-supervised cross-domain few-shot object detection (Semi-CDFSOD) is a practical yet underexplored task that requires adapting detectors to novel domains using scarce support examples alongside abundant unlabeled data. While pseudo-labeling via open-set detectors offers a direct pathway to exploit unlabeled data, it is fundamentally impaired by two intrinsic defects: semantic misclassification, where localized objects are assigned incorrect categories or suppressed as background, and low recall, where the majority of objects are missed entirely. To systematically address these limitations, we propose a training-free pseudo-label calibration framework for Semi-CDFSOD. First, we construct robust class prototypes from the scarce labeled supports and high-confidence pseudo-labels to rectify semantic misclassifications, dynamically recalibrating noisy category assignments while preserving well-localized detections. Second, we introduce a visual self-prompting mechanism that harnesses high-confidence pseudo-labels as prompts for the Segment Anything Model (SAM), generating supplementary proposals to recover undetected objects and dramatically enhance recall. The harmonized pseudo-labels, together with the support samples, are then employed to adapt the final detector. We establish the first comprehensive benchmark for Semi-CDFSOD, spanning six diverse domains. Extensive experiments demonstrate that our method consistently outperforms existing approaches. The code will be released at <https://github.com/xSheep123/Semi-CDFSOD.git>.

Keywords: Cross-Domain Few-Shot Object Learning · Semi-Supervised Learning · Visual Self-Prompting

* Equally contributed first authors. ✉ Corresponding author.

1 Introduction

Cross-domain few-shot object detection (CDFSOD) [6, 18, 19, 25, 26] has emerged as a critical paradigm for adapting detectors to novel domains under extreme label scarcity. Existing approaches, as illustrated in Fig. 1, however, operate under a limiting assumption: that target domains contain nothing beyond the provided support examples. This assumption overlooks a readily available resource—unlabeled data, which is abundant in nearly all real-world deployment scenarios and incurs no annotation cost. The disconnect between problem formulation and practical data availability motivates our investigation of semi-supervised cross-domain few-shot object detection (Semi-CDFSOD), a task that lies at the intersection of domain shift, label scarcity, and unlabeled data abundance, yet remains largely unexplored.

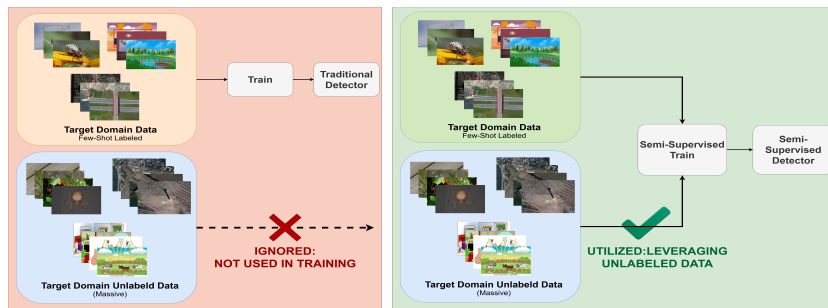


Fig. 1: Comparison of CDFSOD and Semi-CDFSOD paradigms.

The central challenge of Semi-CDFSOD lies in effectively harnessing unlabeled data to bridge the domain gap—which is characterized by distinct differences in both categories and data distributions between the source and target domains—when only a handful of labeled support examples are available per class. Pseudo-labeling, which utilizes a trained detector to predict candidate regions in unlabeled data, offers an intuitive and direct pathway. Recent open-set detectors [5, 12, 20, 28, 35], capable of identifying novel objects beyond their training categories, appear particularly well-suited for this purpose; they can generate pseudo-labels using a detector fine-tuned on few-shot data from the target domain. However, our systematic investigation reveals that this seemingly straightforward approach is fundamentally impaired by two intrinsic defects that arise precisely due to severe domain shift and data scarcity.

The first defect is semantic misclassification. When confronted with unfamiliar domains, open-set detectors, such as MM-Grounding-DINO [35], frequently localize objects accurately but assign them incorrect categories or, worse, suppress them entirely as background. This occurs because the feature distributions of novel domain objects, while distinct from source domain classes, may inadvertently activate similar decision boundaries established by the source domain,

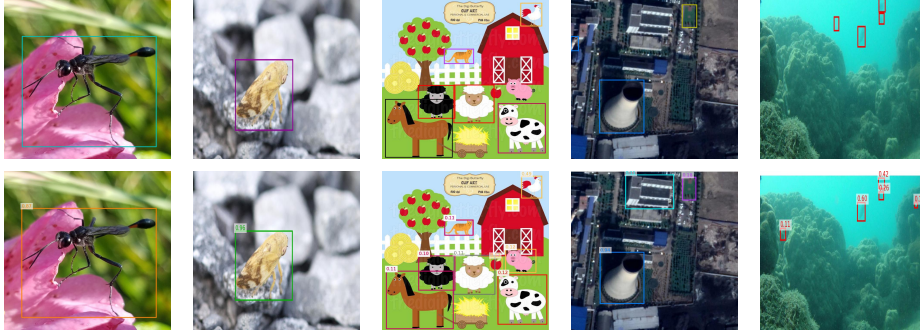


Fig. 2: Qualitative analysis of MM-Grounding-DINO. Representative examples are shown, with ground-truth annotations in the upper row and predicted boxes in the lower row.

leading to confident yet erroneous predictions. The second defect is low recall. The inherent distribution shift often causes feature representations to drift outside the detector’s expected manifold, causing the majority of objects to be missed entirely during proposal generation. These two limitations are not independent; they compound each other, creating a vicious cycle where the few correct detections are drowned out by semantic noise, and the pseudo-labels available for adaptation are both sparse and unreliable. Fig. 2 illustrates this phenomenon: MM-Grounding-DINO localizes certain objects reasonably well yet misclassifies them, hallucinates foregrounds in background regions, and fails to detect valid instances altogether. As shown in Fig. 3, quantitative analysis across datasets confirms that these issues are systematic rather than anecdotal. Consequently, naïvely adopting such predictions as pseudo-labels introduces severely corrupted supervision, fundamentally limiting the utility of unlabeled data.

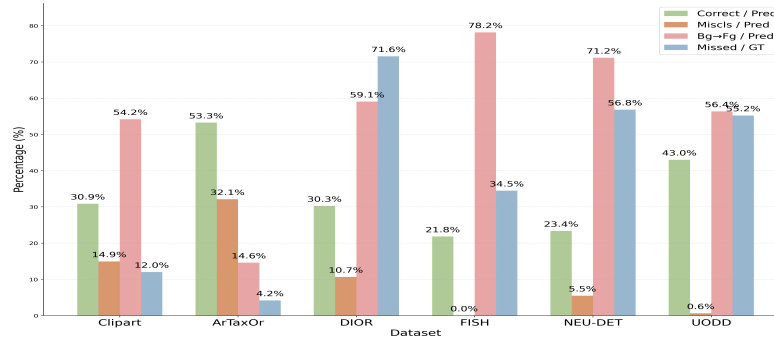


Fig. 3: Quantitative analysis of MM-Grounding-DINO. The figure shows the dataset-level proportions of correct detections, misclassified detections, and missed objects.

Addressing these fundamental limitations requires rethinking the conventional pseudo-labeling pipeline. Rather than accepting noisy pseudo-labels as given and focusing solely on robust adaptation strategies, we argue that the calibration of pseudo-labels themselves must take precedence. The core insight of this work is that the scarce labeled supports, while insufficient for training a detector from scratch, can serve as reliable anchors for calibrating the abundant but noisy pseudo-labels generated by open-set detectors. Furthermore, recent advances in foundation models—particularly the Segment Anything Model (SAM) [3]—offer unprecedented opportunities to recover objects that initial detectors completely miss, effectively serving as a second opinion that complements the detector’s biased perspective.

Building on these insights, we propose a training-free pseudo-label calibration framework that addresses both semantic misclassification and low recall through two complementary mechanisms. First, to rectify noisy category assignments, we design a semantic calibrator that extracts dense features and performs center-weighted ROI [8] aggregation, yielding robust region representations with suppressed background interference. Using these representations, we construct class prototypes from both labeled supports and reliable high-confidence pseudo-labels, then recalibrate pseudo-label categories via cosine-similarity-based prototype matching—preserving accurate localizations while correcting semantic errors. Second, to recover missed objects, we introduce a visual self-prompting mechanism that leverages calibrated high-confidence pseudo-labels as prompts for SAM [3]. Unlike textual prompts, these visual prompts encode both category semantics and target-domain visual characteristics, making them particularly effective under domain shift. The newly discovered instances are subsequently re-calibrated and incorporated as harder examples, enriching the pseudo-label set with objects that were initially undetected. The harmonized pseudo-labels—now semantically accurate and substantially more complete—are finally combined with the original support samples to adapt the detector for the target domain. All operations require no additional training or fine-tuning, relying solely on the frozen backbone, support set, and off-the-shelf SAM.

A critical barrier to progress in Semi-CDFSOD has been the absence of standardized evaluation protocols. Existing work spans disparate datasets, domain shifts, and evaluation metrics, making systematic comparison impossible. To address this gap, we establish the first comprehensive benchmark for Semi-CDFSOD, encompassing six diverse domains with varying degrees of domain gap. For each domain shift, we construct few-shot splits with varying support sizes and provide consistent evaluation protocols to enable fair comparison.

We evaluate our method extensively on this benchmark, comparing against a range of baselines including standard fine-tuning, conventional pseudo-labeling, and state-of-the-art detection methods. The results demonstrate consistent and substantial improvements across all domain shifts and support settings. Notably, our training-free calibration framework outperforms methods that require additional training or complex architectural modifications, highlighting the efficacy

of our core insight: that pseudo-label quality, rather than adaptation strategy complexity, is the primary bottleneck in Semi-CDFSOD.

In summary, this paper makes the following key contributions: (1) We identify and systematically analyze two fundamental limitations of pseudo-labeling for Semi-CDFSOD—semantic misclassification and low recall—revealing how domain shift amplifies these defects and undermines adaptation. (2) We propose a novel training-free calibration framework that addresses both limitations simultaneously. Our approach constructs robust prototypes for semantic rectification and introduces a visual self-prompting mechanism with SAM to recover missed objects, all without additional training. (3) We establish the first comprehensive benchmark for Semi-CDFSOD, spanning six diverse domains with standardized few-shot splits and evaluation protocols, enabling systematic investigation of this underexplored task. Extensive experiments demonstrate that our method consistently outperforms existing approaches.

2 Related Work

2.1 Cross-domain Few-shot Detection

CDFSOD aims to address object detection under both domain shift and limited supervision. Fu et al. [6] introduced a benchmark for this setting, comprising six datasets with pronounced domain disparities to enable systematic evaluation. Building on this, they further explored open-set-detector-based CDFSOD and improved adaptation via fine-tuning, learnable instance features, instance reweighting, and domain prompting. Beyond model adaptation, several studies have attempted to alleviate target-domain data scarcity from a data-centric perspective. Li et al. [18] proposed a training-free compositional image generation framework to produce domain-aligned and detection-friendly samples based on RAG [15]. Pan et al. [26] further observed that although data augmentation can increase sample diversity, it may also introduce instability during detector fine-tuning. Based on this observation, they developed an augmentation-search strategy to dynamically balance the trade-off between diversity and domain relevance. Despite these advances, existing CDFSOD methods still mainly rely on scarce labeled target-domain samples, while the abundant unlabeled target-domain data remain underexplored. This motivates us to investigate a semi-supervised setting for CDFSOD.

2.2 Semi-supervised Few-shot Learning

Semi-supervised few-shot learning (SS-FSL) [27] extends few-shot learning by introducing abundant unlabeled data in addition to a small labeled support set, aiming to alleviate data scarcity and improve generalization under limited supervision. Existing methods mainly improve performance from two perspectives: enhancing class representation and refining pseudo-label quality.

A representative line of work focuses on constructing more robust class prototypes by incorporating unlabeled samples into the support set representation

[2, 14, 21]. By leveraging the feature distribution of unlabeled data, these methods reduce the bias caused by scarce labeled examples and produce more discriminative class representations. Another dominant paradigm is pseudo-labeling [9, 17, 24, 37], where unlabeled samples are assigned estimated labels and then reused as additional supervision to further improve model training. Since the quality of pseudo-labels plays a central role in this process, many methods explicitly address pseudo-label noise and reliability.

Jing et al. [13] measure pseudo-label confidence via certainty entropy and iteratively optimize the classifier using both high-confidence and low-confidence pseudo-labels. Wei et al. [32] instead infer which classes a sample does not belong to from the lowest-confidence predictions, and progressively derive positive pseudo-labels through iterative elimination. Furthermore, [1] leverages the complementarity of different clustering strategies to filter noisy pseudo-labels and improve pseudo-label quality. Xiong et al. [33] first explored semi-supervised few-shot object detection and adopted a teacher-student framework to exploit unlabeled data, alleviating the limited supervision in few-shot object detection.

Although these methods have shown the value of leveraging unlabeled data in few-shot learning, they are not directly applicable to our setting, where pseudo-labeling must jointly address localization and semantic ambiguity under domain shift. This motivates us to develop a more targeted pseudo-label refinement framework for semi-CDFSOD.

3 Method

To address the limitations of semantic misclassification and low recall in pseudo-labels generated by open-set detectors under domain shift, this work presents a training-free pseudo-label calibration framework for Semi-CDFSOD. As illustrated in Fig. 4, the approach operates in two complementary stages: *semantic calibration* to rectify category errors, and *missed object recovery* to retrieve undetected instances—both performed without additional training or fine-tuning.

Given unlabeled target-domain images, initial pseudo-labels are first generated using MM-Grounding-DINO [35], fine-tuned on the limited support examples from the target domain. While these pseudo-labels often exhibit reasonable localization, they suffer from two defects: frequent semantic misclassifications and a substantial number of missed objects. To address both issues, a semantic calibrator is introduced to purify category labels, and a missed-object recovery module is employed to discover additional instances. Recovered candidates are then passed through the same calibrator for semantic verification. Finally, the calibrated and recovered pseudo-labels are merged, non-maximum suppression (NMS) [7] is applied to remove duplicates, and the resulting high-quality pseudo-labels are combined with the original support set for detector adaptation.

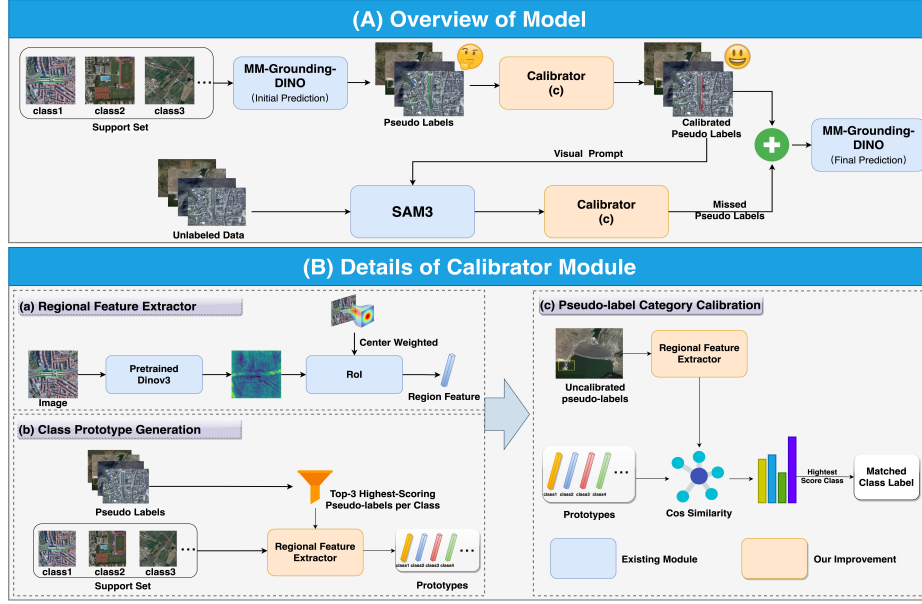


Fig. 4: Overview of our method

3.1 Semantic Calibration

Semantic calibration aims to correct category errors in initial pseudo-labels by matching region features against robust class prototypes, without any additional training. The design is depicted in Fig. 4(B).

Center-Weighted Regional Feature Extraction To obtain discriminative semantic representations for each candidate box, a dense feature map is first extracted using a frozen DINOv3 [30] backbone, which provides strong visual features pre-trained on large-scale data. For a given bounding box, simply averaging features over its ROI inevitably incorporates background noise, especially near the box boundaries. Observing that the object of interest typically occupies the central area of a well-localized box, a center-weighted aggregation strategy is adopted to emphasize informative object regions while suppressing peripheral distractions.

Let O denote the set of patch coordinates inside the ROI, and (c_x, c_y) its center. For each patch $(x, y) \in O$, a center-aware weight is assigned:

$$w_{\text{center}}(x, y) = \exp \left(-\frac{(x - c_x)^2 + (y - c_y)^2}{2\sigma^2} \right), \quad (1)$$

which decays with distance from the center. The weights are normalized across the ROI:

$$\hat{w}(x, y) = \frac{w_{\text{center}}(x, y)}{\sum_{(i, j) \in O} w_{\text{center}}(i, j) + \epsilon}, \quad (2)$$

where ϵ is a small constant for numerical stability. The final region feature $\mathbf{f}_{\text{box}}$ is then computed as the weighted sum of patch features $F(x, y)$:

$$\mathbf{f}_{\text{box}} = \sum_{(x,y) \in O} \hat{w}(x, y) F(x, y). \quad (3)$$

This center-weighted pooling yields a representation that is both semantically precise and robust to imperfect box boundaries.

Class Prototype Construction Reliable semantic calibration hinges on class prototypes that faithfully capture the target-domain appearance of each category. Prototypes are constructed by combining two sources: the few labeled support examples, which provide ground-truth semantics, and high-confidence pseudo-labels, which enrich the prototype with diverse target-domain instances.

For each class c , the top-3 initial pseudo-labels with the highest confidence scores are selected and added to the support samples of that class. All selected regions are passed through the regional feature extractor to obtain a set of features $\{\mathbf{f}_i^c\}_{i=1}^{N_c}$, where N_c is the number of samples used to build the prototype of class c . We then compute the class prototype as

$$\mathbf{p}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbf{f}_i^c. \quad (4)$$

By jointly leveraging the semantic purity of support samples and the statistical richness of high-confidence pseudo-labels, the resulting prototypes provide stable and representative semantic anchors for the target domain.

Prototype-Guided Category Calibration Given the set of class prototypes $\{\mathbf{p}_c\}_{c=1}^C$, the category of each pseudo-labeled region is recalibrated. For a region feature $\mathbf{f}$ extracted from an initial pseudo-label, its cosine similarity with every prototype is computed, and the class corresponding to the highest similarity is assigned:

$$\hat{y} = \arg \max_{c \in \{1, \dots, C\}} \frac{\mathbf{f}^\top \mathbf{p}_c}{\|\mathbf{f}\|_2 \|\mathbf{p}_c\|_2}, \quad (5)$$

where $\hat{y}$ is the calibrated category. This mechanism corrects semantic misclassifications by grounding predictions in the feature space anchored by both support examples and confident pseudo-labels, without any trainable parameters.

3.2 Missed Object Recovery via Visual Self-Prompting

To overcome the low recall of the initial detector, a missed object recovery module is introduced that leverages the calibrated pseudo-labels themselves as visual prompts for SAM. Unlike text prompts, which may suffer from domain ambiguity, visual prompts derived from calibrated detections encode both category

semantics and target-domain visual characteristics, making them particularly effective under domain shift.

Let $\hat{\mathcal{Z}} = \{z_i\}_{i=1}^N$ denote the set of calibrated pseudo-labels, where each z_i includes a bounding box b_i , calibrated category $\hat{y}_i$, and confidence score s_i . $\hat{\mathcal{Z}}$ is partitioned into class-specific subsets $\hat{\mathcal{Z}}^c$. For each class c , the pseudo-label with the highest confidence is selected as the primary visual prompt:

$$z_*^c = \arg \max_{z_i \in \hat{\mathcal{Z}}^c} s_i. \quad (6)$$

This prompt is fed into SAM to search for additional object instances of the same category that were missed by MM-Grounding-DINO.

To further enhance recovery, a class-aware attention guidance mechanism is incorporated. Let $\mathcal{B}^c = \{b_i \mid z_i \in \hat{\mathcal{Z}}^c, z_i \neq z_*^c\}$ be the set of boxes belonging to class c excluding the primary prompt. For each spatial location (x, y) on the feature map, a binary indicator is defined:

$$m^c(x, y) = \begin{cases} 1, & \exists b \in \mathcal{B}^c \text{ s.t. } (x, y) \in b, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Based on this, an enhancement map $e^c(x, y) = \lambda m^c(x, y)$ is constructed, where λ is a fixed coefficient (which we set to 5). This map is added to the box-relative position bias generated from the primary prompt, yielding the final visual cross-attention bias:

$$\mathbf{A}^c = \mathbf{A}_{\text{rpb}}(z_*^c) + \mathbf{E}^c, \quad (8)$$

where $\mathbf{A}_{\text{rpb}}(z_*^c)$ is the position bias induced by the prompt box, and $\mathbf{E}^c$ is the flattened enhancement map derived from $e^c(x, y)$. This guides SAM to attend not only to the region indicated by the most confident prompt, but also to areas where other same-class pseudo-labels reside, thereby improving its ability to discover visually similar but previously missed objects.

For each class c , SAM produces a set of recovered candidates $\mathcal{R}^c = \{r_j^c\}_{j=1}^{M_c}$. After processing all classes, all candidates are collected as $\tilde{\mathcal{Z}} = \bigcup_{c=1}^C \mathcal{R}^c$. These candidates are then passed through the same semantic calibrator to verify their categories and filter out false positives, ensuring that only semantically consistent instances are retained.

Finally, the calibrated pseudo-labels $\hat{\mathcal{Z}}$ and the recovered pseudo-labels $\tilde{\mathcal{Z}}$ are merged to form a comprehensive set. Since the two sets may contain overlapping detections for the same object, NMS is applied to remove duplicates. The resulting high-quality pseudo-labels—now semantically accurate and substantially more complete—are combined with the original support examples to adapt the detector for the target domain.

4 Experiments

4.1 Semi-CDFSOD Benchmark

Problem Formulation. Semi-CDFSOD considers a fully annotated source-domain dataset $\mathcal{D}_S = \{(\mathbf{x}_i^s, y_i^s)\}$ with label space $\mathcal{C}_S$ and distribution P_S , and a target-domain dataset $\mathcal{D}_T$ with label space $\mathcal{C}_T$ and distribution P_T , where $P_S \neq P_T$ and typically $\mathcal{C}_S \cap \mathcal{C}_T = \emptyset$. The target dataset $\mathcal{D}_T$ is partitioned into four disjoint subsets: a few-shot labeled set $\mathcal{D}_{T,f}$, an unlabeled set $\mathcal{D}_{T,u}$, a validation set $\mathcal{D}_{T,v}$, and a test set $\mathcal{D}_{T,t}$. For fair comparison, we follow [6] and adopt the same six cross-domain datasets—ArTaxOr [4], DIOR [16], Clipart1k [10], UODD [11], NEU-DET [31], and FISH [29]. On this basis, we further construct the validation and test splits. Detailed statistics of our benchmark are summarized in Tabs. 1 and 2.

Training and Evaluation Protocol. We follow the standard N -way K -shot setting, where the few-shot set $\mathcal{D}_{T,f}$ provides exactly K annotated bounding boxes for each of the N target categories, with $N = |\mathcal{C}_T|$. The detector is first pre-trained on the source domain $\mathcal{D}_S$, and then adapted to the target domain using both the labeled few-shot set $\mathcal{D}_{T,f}$ and the unlabeled set $\mathcal{D}_{T,u}$. Model selection is performed on the validation set $\mathcal{D}_{T,v}$, and final performance is reported on the test set $\mathcal{D}_{T,t}$.

Table 1: Dataset split statistics of our Semi-CDFSOD benchmark. We report the number of images and instances for each subset.

Dataset	$ \mathcal{C}_T $	Unlabeled set $\mathcal{D}_{T,u}$		Validation set $\mathcal{D}_{T,v}$		Test set $\mathcal{D}_{T,t}$	
		#Images	#Boxes	#Images	#Boxes	#Images	#Boxes
ArTaxOr	7	205	7×40	155	7×30	1383	1628
FISH	1	13	1×40	7	1×30	90	3029
DIOR	20	252	20×40	182	20×30	5000	28810
NEU-DET	6	121	6×40	70	6×30	360	834
UODD	3	21	3×40	16	3×30	506	3218
Clipart1k	20	46	162	30	109	461	1473

Table 2: Per-class statistics on Clipart1k.

Split	C4	C5	C6	C7	C8	C10	C11	C12	C14	C17	C18	C19	C20
$\mathcal{D}_{T,u}$ (Semi-sup.)	10	10	10	6	0	10	9	4	4	9	0	10	10
$\mathcal{D}_{T,v}$ (Val.)	6	7	3	6	0	7	2	1	1	1	0	5	0

Note. Classes with 10 instances in both $\mathcal{D}_{T,u}$ and $\mathcal{D}_{T,v}$ are omitted.

4.2 Implementation Details.

Our baseline is the pretrained MM -GroundingDINO [35] with a Swin-B [22] backbone, initialized with weights from LLMDet [5] to leverage its superior open-vocabulary capabilities. For the feature extraction module, we utilize DI-NOv3 [30] (ViT-L/16). The detector is fine-tuned for 500 iterations using the AdamW optimizer [23]. The base learning rate is set to 2×10^{-5} (1×10^{-4} specifically for ArTaxOr and DIOR), with the backbone learning rate scaled by a factor of 0.1.

4.3 Main Results

Tabs. 3 to 5 summarize the detection performance (mAP) on six cross-domain datasets under the 1/5/10-shot settings. The best and second-best results in each case are highlighted in bold and underlined, respectively.

For comprehensive comparison, we group existing methods into three categories: (1) traditional CDFSOD methods (e.g., Detic, DE-ViT, CD-ViTO); (2) pseudo-label filtering strategies based on confidence thresholds; and (3) pseudo-label generation methods leveraging vision foundation models. Across most datasets and shot settings, our method achieves superior performance, attaining the highest average mAP of 35.9, 44.1, and 46.7 under 1, 5, and 10 shots, respectively.

Table 3: Results (mAP) under the Semi-CDFSOD 1-shot setting.

Method	ArTaxOr	Clipart1k	DeepFish	DIOR	NEU-DET	UODD	Avg.
Detic-FT [36]	3.2	15.1	9.0	4.1	3.8	4.2	6.6
DE-ViT-FT [34]	10.5	13.0	19.3	14.7	0.6	2.4	10.1
CD-ViTO [6]	21.0	17.7	20.3	17.8	3.6	3.1	13.9
MM-Grounding-DINO [35]	44.0	49.6	39.0	18.6	13.9	<u>22.2</u>	31.2
Threshold-based PL Filtering(thr=0)	9.9	42.9	38.2	4.3	3.2	19.6	19.7
Threshold-based PL Filtering(thr=0.5)	44.9	51.7	40.1	18.9	15.4	19.6	31.8
Threshold-based PL Filtering(thr=0.9)	<u>52.2</u>	<u>50.3</u>	38.2	18.4	15.0	22.5	<u>32.8</u>
SAM3 [3] w/ Text Prompts	39.2	48.9	37.6	17.9	7.6	21.6	28.8
SAM3 [3] w/ Visual Prompts	47.6	46.0	<u>41.6</u>	<u>20.1</u>	<u>15.6</u>	21.6	32.1
SAM3 [3] w/ Text & Visual Prompts	51.2	49.3	38.3	18.5	9.9	20.7	31.3
Ours	56.2	50.1	46.4	23.7	16.8	22.0	35.9

Experimental results demonstrate that incorporating semi-supervised mechanisms consistently outperforms traditional CDFSOD methods that rely solely on fine-tuning with limited labeled samples. Under domain shift, fine-tuning with few annotated examples is prone to overfitting, whereas leveraging unlabeled target-domain data mitigates this issue and improves cross-domain generalization.

We further evaluate the confidence threshold-based pseudo-label filtering strategy. This approach proves highly sensitive to threshold selection. A low threshold (thr=0) introduces excessive noisy pseudo-labels, causing substantial

Table 4: Results (mAP) under the Semi-CDFSOD 5-shot setting.

Method	ArTaxOr	Clipart1k	DeepFish	DIOR	NEU-DET	UODD	Avg.
Detic-FT [36]	8.7	20.2	14.3	12.1	14.1	10.4	13.3
DE-ViT-FT [34]	38.0	38.1	21.2	23.4	7.8	5.0	22.3
CD-ViTO [6]	47.9	41.1	22.3	26.9	11.4	6.8	26.1
MM-Grounding-DINO [35]	67.6	55.6	41.0	<u>29.6</u>	22.2	27.8	40.6
Threshold-based PL Filtering(thr=0)	14.0	39.8	39.0	7.9	4.9	20.4	21.0
Threshold-based PL Filtering(thr=0.5)	<u>76.2</u>	55.3	<u>42.2</u>	27.3	<u>22.3</u>	24.2	<u>41.3</u>
Threshold-based PL Filtering(thr=0.9)	71.5	56.4	40.9	28.9	21.0	<u>28.0</u>	41.1
SAM3 [3] w/ Text Prompts	51.3	50.6	37.1	24.8	13.1	26.1	33.8
SAM3 [3] w/ Visual Prompts	73.3	53.5	38.6	28.7	21.1	27.7	40.5
SAM3 [3] w/ Text & Visual Prompts	70.8	51.2	37.9	27.5	12.7	27.5	37.9
Ours	78.6	<u>55.8</u>	43.3	32.5	23.8	30.3	44.1

Table 5: Results (mAP) under the Semi-CDFSOD 10-shot setting.

Method	ArTaxOr	Clipart1k	DeepFish	DIOR	NEU-DET	UODD	Avg.
Detic-FT [36]	12.0	22.3	17.9	15.4	16.8	14.4	16.5
DE-ViT-FT [34]	49.2	40.8	21.3	25.6	8.8	5.4	25.2
CD-ViTO [6]	60.5	44.3	22.3	30.8	12.8	7.0	29.6
MM-Grounding-DINO [35]	76.7	57.7	<u>44.4</u>	35.7	24.3	<u>31.7</u>	45.1
Threshold-based PL Filtering(thr=0)	15.9	51.1	39.3	10.9	4.2	20.4	23.6
Threshold-based PL Filtering(thr=0.5)	<u>80.5</u>	57.1	44.2	37.3	23.3	29.5	<u>45.3</u>
Threshold-based PL Filtering(thr=0.9)	74.2	<u>57.4</u>	44.0	<u>36.0</u>	<u>24.6</u>	30.7	44.5
SAM3 [3] w/ Text Prompts	63.0	53.6	38.5	27.6	14.8	30.0	37.9
SAM3 [3] w/ Visual Prompts	80.1	56.3	41.6	34.7	21.5	30.2	44.1
SAM3 [3] w/ Text & Visual Prompts	76.1	53.6	39.6	28.3	14.0	28.7	40.1
Ours	82.4	<u>57.4</u>	44.8	37.3	25.9	32.2	46.7

performance degradation (e.g., 9.9 mAP on 1-shot ArTaxOr). A high threshold (thr=0.9) reduces noise but retains mostly easy examples, limiting the exploitation of objects in unlabeled data. These results indicate that discarding low-confidence pseudo-labels via a fixed threshold cannot adequately address the two core challenges identified in our analysis—semantic misclassification and low recall.

Comparing pseudo-label generation using SAM-based models, we observe clear performance differences across prompt strategies. Visual prompts consistently outperform text prompts, suggesting that visual cues are more effective for object discovery under domain shift. However, when prompts are derived from noisy pseudo-labels, they may amplify model bias, creating a “prompt noise → output noise” loop. Our method mitigates this by calibrating categories prior to visual prompting, suppressing noise propagation and improving both pseudo-label quality and training stability.

4.4 Ablation Study

To verify the effectiveness of the category calibration module and the visual self-prompting module, we conduct progressive ablation studies on the DIOR and NEU-DET datasets. Specifically, we first apply the category calibration module to refine the initial pseudo-labels generated by the baseline, and then

further introduce the visual self-prompting module to recover potentially missed objects. The detailed results are reported in Tab. 6.

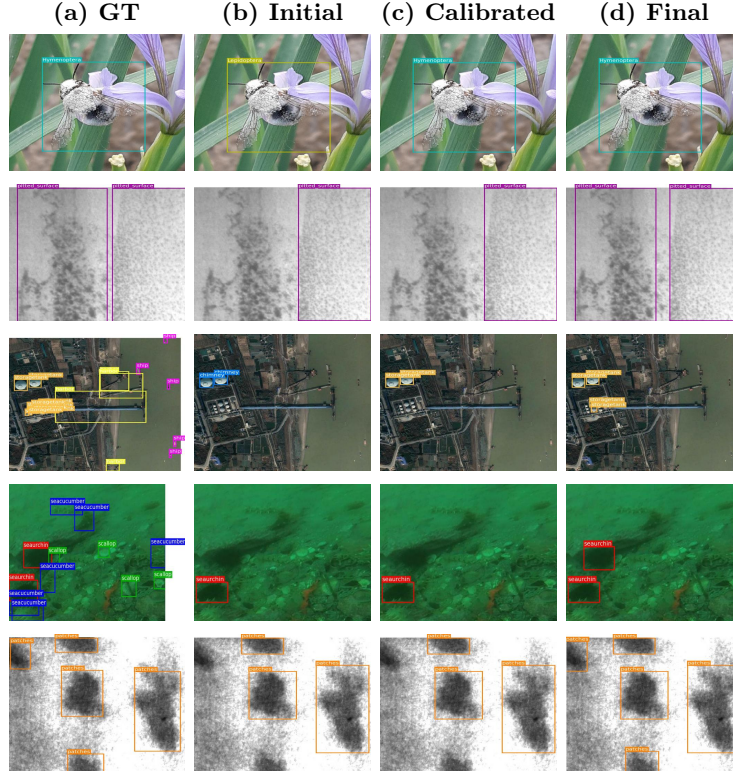


Fig. 5: Visual comparison of predictions at different stages across multiple target domains. From left to right: ground truth, initial predictions, calibrated predictions, and final predictions.

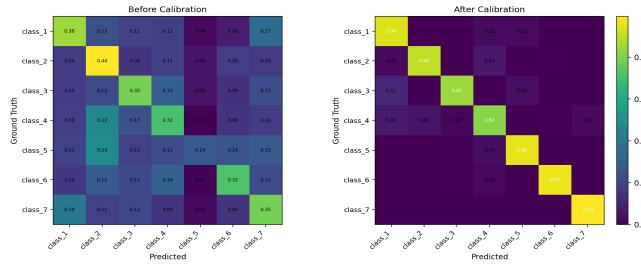
The results in Tab. 6 show that both the calibrator module and the visual self-prompting module bring consistent performance gains. After introducing the calibrator, the baseline achieves clear improvements on both DIOR and NEU-DET under the 1/5/10-shot settings; further incorporating visual self-prompting leads to additional gains and yields the best results on both datasets. This demonstrates that the two modules are complementary, jointly improving detection performance through the calibrator and missed-object completion.

We first provide qualitative comparisons to intuitively illustrate the effects of our method. As shown in Fig. 5, the initial pseudo-labels from MM-Grounding-DINO exhibit both semantic misclassifications and missed detections. After applying our calibration framework, the calibrated pseudo-labels exhibit corrected

Table 6: Ablation study on DIOR and NEU-DET.

Method	DIOR			NEU-DET		
	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
Baseline	18.6	29.6	35.7	13.9	22.2	24.3
+ calibrator	20.6	32.1	36.0	16.5	23.5	25.6
+ calibrator & visual self-prompting (Ours)	23.7	32.5	37.3	16.8	23.8	25.9

categories, while the recovery module complements previously missed objects, yielding a more complete and accurate set.

**Fig. 6:** Comparison of Confusion Matrices Before and After Calibration

To further quantify the improvement in semantic accuracy, Fig. 6 presents the confusion matrices before and after category calibration. The diagonal responses are enhanced while off-diagonal responses are reduced, indicating that inter-class confusion is effectively suppressed. This confirms that the category calibration module successfully corrects the semantic noise in pseudo-labels, addressing the misclassification defect identified in our analysis.

5 Conclusion

We present a new Semi-CDFSOD benchmark that incorporates low-cost unlabeled target-domain data to mitigate annotation scarcity. Under this setting, we observe that MM-Grounding-DINO—even after few-shot fine-tuning—still exhibits two fundamental limitations in pseudo-label generation: semantic misclassification and low recall, both of which are exacerbated by domain shift. To address these issues, we propose a training-free pseudo-label calibration framework. It performs semantic calibration via robust region representations and class prototypes constructed from both support examples and high-confidence pseudo-labels, and recovers missed objects by leveraging calibrated pseudo-labels as visual prompts for SAM. Extensive experiments demonstrate that our method consistently yields stable and significant gains across diverse target domains without any additional training.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62472359, 62372379), in part by the Xi'an's Key Industrial Chain Core Technology Breakthrough Project: AI Core Technology Breakthrough under Grant 24ZDCYJSGG0003.

References

1. Al-Habib, M., Zhang, Z., Noman, A., Al-Qatf, M., Abdulaziz, M.: Semi-supervised few-shot learning via ensemble-based adaptive clustering-aware pseudo-labeling. *Applied Soft Computing* p. 114225 (2025)
2. Ayyad, A., Li, Y., Muaz, R., Albarqouni, S., Elhoseiny, M.: Semi-supervised few-shot learning with prototypical random walks. In: *AAAI Workshop on Meta-Learning and MetaDL Challenge*. pp. 45–57. PMLR (2021)
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025)
4. Drange, G.: Arthropod taxonomy orders object detection dataset (2020). <https://doi.org/10.34740/KAGGLE/DSV/1240192>, <https://www.kaggle.com/dsv/1240192>
5. Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., Zheng, W.S.: Llm-det: Learning strong open-vocabulary object detectors under the supervision of large language models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 14987–14997 (2025)
6. Fu, Y., Wang, Y., Pan, Y., Huai, L., Qiu, X., Shangguan, Z., Liu, T., Fu, Y., Van Gool, L., Jiang, X.: Cross-domain few-shot object detection via enhanced open-set object detector. In: *European Conference on Computer Vision*. pp. 247–264. Springer (2024)
7. Girshick, R.: Fast r-cnn. In: *Proceedings of the IEEE international conference on computer vision*. pp. 1440–1448 (2015)
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 580–587 (2014)
9. Huang, K., Geng, J., Jiang, W., Deng, X., Xu, Z.: Pseudo-loss confidence metric for semi-supervised few-shot learning. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 8671–8680 (2021)
10. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5001–5009 (2018)
11. Jiang, L., Wang, Y., Jia, Q., Xu, S., Liu, Y., Fan, X., Li, H., Liu, R., Xue, X., Wang, R.: Underwater species detection using channel sharpening attention. In: *Proceedings of the 29th ACM International Conference on Multimedia*. pp. 4259–4267 (2021)
12. Jiang, Q., Li, F., Zeng, Z., Ren, T., Liu, S., Zhang, L.: T-rex2: Towards generic object detection via text-visual prompt synergy. In: *European Conference on Computer Vision*. pp. 38–57. Springer (2024)
13. Jing, K., Ma, H., Zhang, C., Wen, L., Zhang, Z.: Recursive confidence training for pseudo-labeling calibration in semi-supervised few-shot learning. *IEEE Transactions on Image Processing* (2025)

14. Lan, X., McDonagh, S., Gong, S., Wang, J., Li, Z., Parisot, S.: Fewmatch: Dynamic prototype refinement for semi-supervised few-shot learning (2020)
15. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **33**, 9459–9474 (2020)
16. Li, K., Wan, G., Cheng, G., Meng, L., Han, J.: Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing* **159**, 296–307 (2020)
17. Li, P., Wu, G., Gong, S., Lan, X.: Semi-supervised few-shot learning with pseudo label refinement. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
18. Li, Y., Qiu, X., Fu, Y., Chen, J., Qian, T., Zheng, X., Paudel, D.P., Fu, Y., Huang, X., Van Gool, L., et al.: Domain-rag: Retrieval-guided compositional image generation for cross-domain few-shot object detection. *arXiv preprint arXiv:2506.05872* (2025)
19. Liu, C., Xiang, X., Duan, Z., Li, W., Fan, Q., Gao, Y.: Don’t need retraining: A mixture of detr and vision foundation models for cross-domain few-shot object detection. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
20. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
21. Liu, Y., Wan, H., Shi, D.: Semi-supervised few-shot learning using critical-descriptor prototypes. *Neurocomputing* p. 130769 (2025)
22. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
24. Ma, C., Dong, W., Xu, C.: Tenet: Beyond pseudo-labeling for semi-supervised few-shot learning. *Machine Intelligence Research* **22**(3), 511–523 (2025)
25. Meng, B., Zhang, X., Li, P., Wu, Z., Li, Y., Zhao, W., Yu, B., Shen, H.L.: Cdformer: Cross-domain few-shot object detection transformer against feature confusion. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025)
26. Pan, J., Liu, Y., He, X., Peng, L., Li, J., Sun, Y., Huang, X.: Enhance then search: An augmentation-search strategy with foundation models for cross-domain few-shot object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1548–1556 (2025)
27. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676* (2018)
28. Ren, T., Chen, Y., Jiang, Q., Zeng, Z., Xiong, Y., Liu, W., Ma, Z., Shen, J., Gao, Y., Jiang, X., et al.: Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347* (2024)
29. Saleh, A., Laradji, I.H., Konovalov, D.A., Bradley, M., Vazquez, D., Sheaves, M.: A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific reports* **10**(1), 14671 (2020)

30. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khali-dov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
31. Song, K., Yan, Y.: A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science* **285**, 858–864 (2013)
32. Wei, X.S., Xu, H.Y., Yang, Z., Duan, C.L., Peng, Y.: Negatives make a positive: An embarrassingly simple approach to semi-supervised few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(4), 2091–2103 (2023)
33. Xiong, W., Cui, Y., Liu, L.: Semi-supervised few-shot object detection with a teacher-student network. In: *BMVC*. p. 290 (2021)
34. Zhang, X., Liu, Y., Wang, Y., Boularias, A.: Detect everything with few examples. In: *8th Annual Conference on Robot Learning*
35. Zhao, X., Chen, Y., Xu, S., Li, X., Wang, X., Li, Y., Huang, H.: An open and comprehensive pipeline for unified object grounding and detection. arXiv preprint arXiv:2401.02361 (2024)
36. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: *European conference on computer vision*. pp. 350–368. Springer (2022)
37. Zhu, L., Ke, Z., Lau, R.: Towards self-adaptive pseudo-label filtering for semi-supervised learning. arXiv preprint arXiv:2309.09774 (2023)