Open-World Objectness Modeling Unifies Novel Object Detection

Supplementary Material

Authors:
Shan Zhang, Yao Ni, Jinhao Du, Yuan Xue, Philip Torr, Piotr Koniusz, Anton van den Hengel.

In this supplementary material, we validate the soundness of the dynamic Gaussian prior both theoretically and experimentally (§A), evaluate performance using WI and A-OSE metrics (§B), present incremental learning experiments and additional ablation studies (§C), visualize the box detection results on novel classes (§D), and include a detailed discussion of related work (§E).

A. Soundness of dynamic Gaussian prior

In Eq. 3, we propose a dynamic Gaussian prior for $\mathcal{L}_{KL}$ to resolve the issue of losing diversity in low-data regimes. To demonstrate the importance of the dynamic Gaussian prior in building a robust objectness model for reducing the cross-entropy between class and objectness variables (formulated as pseudo-labeling), we provide theoretical analysis and experimental verification.

A.1. Theoretical Analysis

Starting with a key lemma to lay the foundation for our proof, we delve into the probabilistic objectness model, highlighting the risks of losing diversity, especially in low-data regimes. Our analysis then shows that dynamic Guassian prior introduced into our $\mathcal{L}_{KL}$ effectively mitigates this issue.

Lemma 1 *Given random variables X_1, X_2 from two distributions, we sample data at a ratio of α to $1-\alpha$ to create X. With μ_1, μ_2 and σ_1^2, σ_2^2 representing the mean and variance of X_1, X_2 respectively, the mean and variance of of X are:*

$$\begin{cases} E[X] = \alpha E[X_1] + (1-\alpha)E[X_2] = \alpha\mu_1 + (1-\alpha)\mu_2 \\ D[X] = \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + \alpha(1-\alpha)(\mu_1-\mu_2)^2 \end{cases} \tag{A.1}$$

***Proof**: Below we provide the proof for the variance:*

$$\begin{aligned} D[X] =& E[(X-E[X])^2] = E\big[X - \big(\alpha\mu_1 + (1-\alpha)\mu_2\big)^2\big] \\ =& \alpha E[(X_1 - \mu_1 + \mu_1 - (\alpha\mu_1 + (1-\alpha)\mu_2))^2] \\ &+ (1-\alpha)E[(X_2 - \mu_2 + \mu_2 - (\alpha\mu_1 + (1-\alpha)\mu_2))^2] \\ =& \alpha E[(X_1-\mu_1)^2] + \alpha[(1-\alpha)(\mu_1-\mu_2)]^2 \\ &+ (1-\alpha)E[(X_2-\mu_2)^2] + (1-\alpha)[\alpha(\mu_1-\mu_2)]^2 \\ =& \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + \alpha(1-\alpha)(\mu_1-\mu_2)^2 \end{aligned}$$

A.1.1. Issue of losing diversity

Theorem 1 *Consider a query embedding $\check{\boldsymbol{q}}_t$ from known classes at iteration step t, following $\mathcal{N}(\check{\boldsymbol{\mu}}_t, \check{\boldsymbol{\sigma}}_t^2)$, and a query embedding $\tilde{\boldsymbol{q}}$ from unknown objects or backgrounds, following $\mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}^2)$. Combining two distributions in a α to $1-\alpha$ data ratio, we model them as a single distribution $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2)$, simplifying our approach by using $\boldsymbol{\sigma}_t^2$ instead of a covariance matrix. Keeping $\tilde{\boldsymbol{q}}$ fixed, we alternate between estimating the mean/variance of this unified distribution:*

$$\begin{cases} \boldsymbol{\mu}_{t+1} = \alpha\check{\boldsymbol{\mu}}_t + (1-\alpha)\tilde{\boldsymbol{\mu}} \\ \boldsymbol{\sigma}_{t+1}^2 = \alpha\check{\boldsymbol{\sigma}}_t^2 + (1-\alpha)\tilde{\boldsymbol{\sigma}}^2 + \alpha(1-\alpha)(\check{\boldsymbol{\mu}}_t - \tilde{\boldsymbol{\mu}})^2 \end{cases} \tag{A.2}$$

and maximization likelihood over queries from known classes by minimizing:

$$\min_{\check{\boldsymbol{q}}_t} \Big(\frac{\check{\boldsymbol{q}}_t - \boldsymbol{\mu}_t}{\boldsymbol{\sigma}_t}\Big)^2 \tag{A.3}$$

This alternating process between statistical estimation and likelihood maximization leads to convergence as $t \to \infty$:

$$\begin{cases} \check{\boldsymbol{q}}_{t\to\infty} = \tilde{\boldsymbol{\mu}} \\ \boldsymbol{\mu}_{t\to\infty} = \tilde{\boldsymbol{\mu}}, \quad \boldsymbol{\sigma}^2_{t\to\infty} = (1-\alpha)\tilde{\boldsymbol{\sigma}}^2 \\ \check{\boldsymbol{\mu}}_{t\to\infty} = \tilde{\boldsymbol{\mu}}, \quad \check{\boldsymbol{\sigma}}^2_{t\to\infty} = \mathbf{0} \end{cases} \tag{A.4}$$

Remark: Theorem 1 reveals that alternating optimization reduces the diversity of query embeddings for known classes. As $t \to \infty$, during the distribution estimation step, the mean value $\boldsymbol{\mu}$ approaches $\tilde{\boldsymbol{\mu}}$ and *its variance $\boldsymbol{\sigma}$ decreases by* $1-\alpha$. During the maximization step, known queries $\check{\boldsymbol{q}}$ concentrate densely around $\boldsymbol{\mu}$, with *their variance approaching zero*. In other words, the high density of known queries and reduced variance diminish diversity, leading to bias toward known classes. This issue becomes more pronounced in low-data regimes with limited training data, where diversity diminishes rapidly at small t.

Proof: Eq. A.2 follows directly from Lemma 1. We proceed to prove Eq. A.4. By minimizing Eq. A.3, we find that the optimal scenario occurs when all $\check{\boldsymbol{q}}_t = \boldsymbol{\mu}_t$, implying $\check{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_t$ and $\check{\boldsymbol{\sigma}}_t = \mathbf{0}$. Denote $\check{\boldsymbol{\mu}}_0$ as the initial mean of embeddings of known classes, then:

$$\begin{aligned} \boldsymbol{\mu}_1 &= \alpha\check{\boldsymbol{\mu}}_0 + (1-\alpha)\tilde{\boldsymbol{\mu}} \\ \boldsymbol{\mu}_2 &= \alpha^2\check{\boldsymbol{\mu}}_0 + \alpha(1-\alpha)\tilde{\boldsymbol{\mu}} + (1-\alpha)\tilde{\boldsymbol{\mu}} = \alpha^2\check{\boldsymbol{\mu}}_0 + (1-\alpha^2)\tilde{\boldsymbol{\mu}} \\ \boldsymbol{\mu}_t &= \alpha^t\check{\boldsymbol{\mu}}_0 + (1-\alpha^t)\tilde{\boldsymbol{\mu}} \end{aligned}$$

Given $0 < \alpha < 1$, as $t \to \infty$, $\alpha^t \to 0$. Thus $\boldsymbol{\mu}_{t\to\infty} = \tilde{\boldsymbol{\mu}}$, $\check{\boldsymbol{\mu}}_{t\to\infty} = \tilde{\boldsymbol{\mu}}$, and $\check{\boldsymbol{\sigma}}^2_{t\to\infty} = \mathbf{0}$, resulting in $\boldsymbol{\sigma}^2_{t\to\infty} = (1-\alpha)\tilde{\boldsymbol{\sigma}}^2$. This completes the proof.

A.1.2. Dynamic Gaussian prior perseveres diversity

As noted in the main paper, the KL divergence between the estimated objectness posterior and a static normal prior mitigates overfitting and prevents latent space collapse. However, it suffers from logarithmic divergence in low-data scenarios. To address this, we inject Gaussian noise into the estimated objectness posterior, forming a dynamic Gaussian prior as a surrogate. Herein, we demonstrate that the dynamic Gaussian prior performs a similar function to the static normal prior, mitigating overfitting and preserving the diversity of latent embeddings, with its divergence dynamically adjusted by the parameter β. Below, we analyze its effect on the optimization of latent query embeddings.

Theorem 2 *Using notations from Theorem 1, we introducing a disturbance $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\beta}^2)$ to estimated objectness prior during likelihood maximization, represented as:*

$$\min_{\check{\boldsymbol{q}}_t} \left(\frac{\check{\boldsymbol{q}}_t - \boldsymbol{\epsilon} - \boldsymbol{\mu}_t}{\boldsymbol{\sigma}_t} \right)^2 \tag{A.5}$$

As $t \to \infty$, the alternating process between likelihood maximization and statistical estimation yields:

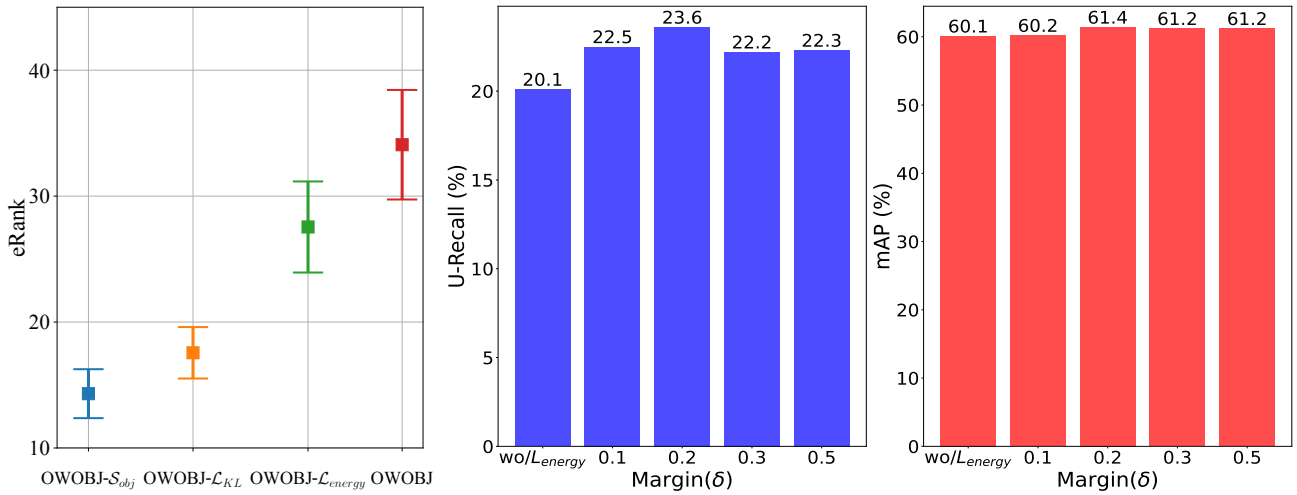
$$\begin{cases} \check{\boldsymbol{q}}_{t\to\infty} = \tilde{\boldsymbol{\mu}} + \boldsymbol{\epsilon}, \textit{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\beta}^2) \\ \boldsymbol{\mu}_{t\to\infty} = \tilde{\boldsymbol{\mu}}, \quad \boldsymbol{\sigma}^2_{t\to\infty} = (1-\alpha)\tilde{\boldsymbol{\sigma}}^2 + \alpha\boldsymbol{\beta}^2 \\ \check{\boldsymbol{\mu}}_{t\to\infty} = \tilde{\boldsymbol{\mu}}, \quad \check{\boldsymbol{\sigma}}^2_{t\to\infty} = \boldsymbol{\beta}^2 \end{cases} \tag{A.6}$$

Remark: Compared with Theorem 1, Theorem 2 reveals that during likelihood maximization, adding noise introduces a relaxation factor $\boldsymbol{\epsilon}$ that enables the embeddings from known classes to retain some degree of diversity and discriminability. This relaxation, coupled with increased variance $\boldsymbol{\sigma}$, reduces the tendency to concentrate on known classes, allowing higher scores for potential unknown objects than without noise.

Proof: Following a derivation similar to Theorem 1, we find that the optimal outcome by minimizing Eq. A.5, is $\check{\boldsymbol{q}}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\beta}^2)$. This leads to

$$\check{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_t, \check{\boldsymbol{\sigma}}^2_t = \boldsymbol{\beta}^2 \tag{A.7}$$

As with Theorem 1, when $t \to \infty$, $\boldsymbol{\mu}_{t\to\infty} = \tilde{\boldsymbol{\mu}}$. Using Lemma 1, we deduce $\boldsymbol{\sigma}^2_{t\to\infty} = (1-\alpha)\tilde{\boldsymbol{\sigma}}^2 + \alpha\boldsymbol{\beta}^2$, thus completing the proof.



(a) (b) (c)

Figure 5. (a) Effective rank (eRank) [40] for each method, averaged over training on Taks1. OWOBJ−$\mathcal{S}_{obj}$: removes objectness score and replaces pseudo-labels from the probabilistic objectness model with random soft labels; OWOBJ−$\mathcal{L}_{KL}$: removes $\mathcal{L}_{KL}$; OWOBJ−$\mathcal{L}_{energy}$: removes $\mathcal{L}_{energy}$. (b) U-Recall (%) for unknown detection and (c) mAP (%) for known detection of PROB+OWOBJ w.r.t. vaying margins (δ) in $\mathcal{L}_{energy}$ for Task1 on M-OWODB dataset.

A.2. Ablation studies for dynamic Guassian prior

To validate our analysis and the efficacy of dynamic Gaussian prior, we analyze the effective rank (eRank) [40], a measure of feature diversity, where a higher eRank indicates better preservation of information or feature diversity, and thus, potentially superior representation. We compare the eRank across different methods: OWOBJ−$\mathcal{S}_{obj}$, OWOBJ−$\mathcal{L}_{KL}$, OWOBJ−$\mathcal{L}_{energy}$, and our final method OWOBJ. As shown in Fig. 5a, the eRank of variants OWOBJ−$\mathcal{S}_{obj}$, which removes objectness score and using randomly generated soft labels for pseudo-labeling, and OWOBJ−$\mathcal{L}_{KL}$, which removes $\mathcal{L}_{KL}$, are significantly lower than that of OWOBJ. This supports the assertion in Theorem 2 that noise addition enhances feature diversity. Moreover, OWOBJ achieves the highest eRank, confirming its effectiveness in addressing the issue of losing diversity and producing meaningful, diverse feature representations.

A.3. Ablation studies for margin δ

To assess the impact of margin δ on performance, we varied δ and plotted the results on Task 1 of OWOD, using U-Recall for unknown detection and mAP for known objects, as shown in Fig. 5b and 5c. We can see that without $\mathcal{L}_{energy}$, both U-Recall and mAP decreased by 2.5% and 1.3%, respectively. As δ increased to 0.2, the performance for known objects saturated, showing limited sensitivity to the margin δ. However, the U-Recall for unknown objects begins to decline, suggesting that larger margins introduce uncertainty by pushing unknown detections into low-probability regions. Thus, $\delta = 0.2$ achieves the best balance between known and unknown object detection.

B. Evaluation using WI and A-OSE Metrics

In §5.1, we discuss additional evaluation metrics that shed light on the performance of different open-world object detection (OWOD) methods. Table 5 compares these methods on the M-OWODB dataset [20], focusing on unknown recall (U-recall), wilderness impact (WI), and absolute open-set error (A-OSE). Among these metrics, U-recall accesses the effectiveness of OWOD models in detecting unknown objects, serving as an indicator of their proficiency of identifying unlabeled objects. WI and A-OSE, in contrast, measure the model's susceptibility to confuse unknown instances with known classes. Specifically, WI quantifies the impact of unknown detections on the model's precision, noting that as U-recall increases, the prominence of unknown object precision can lead to a rise in WI, even when the models exhibit similar unknown object precision. Therefore, WI tends to escalate with U-recall. On the other hand, A-OSE measures the total count of unknown instances mistakenly identified as known classes, with less sensitivity to changes in U-recall.

Upon reviewing Table 5, it is evident that PROB+OWOBJ surpasses all other OWOD techniques in U-recall while simultaneously maintaining lower A-OSE values. This indicates that our method excels at detecting unknown objects while

reducing the misclassification of unknown instances as known classes.

Table 5. State-of-the-art comparison for unknown object confusion on M-OWODB using wilderness impact (WI), absolute open set error (A-OSE) and unknown class recall (U-Recall, which quantifies a model's ability to retrieve unknown object instances). PROB+OWOBJ outperforms PROB in WI, A-OSE, and U-Recall across multiple tasks, indicating reduced misclassification of unknowns as background and improved detection of unknown instances. Note that WI, A-OSE and U-Recall are not applied to task 4, where all 80 classes are known, and are therefore excluded.

Task IDs (→)	**Task 1**			**Task 2**			**Task 3**		
	U-Recall (↑)	WI (↓)	A-OSE (↓)	U-Recall (↑)	WI (↓)	A-OSE (↓)	U-Recall (↑)	WI (↓)	A-OSE (↓)
ORE − EBUI [20]	4.9	0.0621	10459	2.9	0.282	10445	3.9	0.0211	7990
2B-OCD [51]	12.1	0.0481	-	9.4	0.160	-	11.6	0.0137	-
OW-DETR [17]	7.5	0.0571	10240	6.2	0.0278	8441	5.7	0.0156	6803
OCPL [34]	8.3	0.0413	5670	7.6	0.0220	5690	11.9	0.0162	5166
PROB [71]	19.4	0.0569	5195	17.4	0.0344	6452	19.6	0.0151	2641
PROB+OWOBJ	23.6	**0.03951**	**3102**	**23.8**	**0.0215**	**4032**	**25.1**	**0.0085**	**821**

C. More Experiments

C.1. Incremental Object Detection

Previous studies [17, 71] showed that incremental object detection (iOD) performs well at detecting unknown objects. This success is attributed to its ability to reduce confusion between unknown objects, known classes and background, allowing the detector to progressively learn new object classes as foreground instances.

We evaluate our method on incremental object detection tasks, with results presented in Tab. 6. The table compares PROB+OWOBJ with existing approaches on PASCAL VOC 2007, using the evaluation protocols from [17, 51]. In each evaluation scenario, the model is first trained on a subset of object classes (10/15/19), followed by the incremental introduction of additional classes (10/5/1). Our method achieves final mAP scores of 69.9, 73.3, and 75.8, surpassing the previous state-of-the-art, PROB, which achieves scores of 66.5, 70.1, and 72.6, respectively.

C.2. Add-on with other methods

We integrate our OWOBJ into MEPU-FS [14] and CAT [33]. For MEPU-FS, we substitute its REW with our f^{pr}_{obj} for scoring unknown proposals and replace Eq. 5 in MEPU-FS with our Eq. 6 to assign objectness scores for unknown proposals. As shown in Tab. 7, we achieve a 2.3%-3.1% improvement in U-Recall and a 1.8%-2.4% increase in known mAP across Tasks 1-4 on S-OWODB. Building upon CAT (with cascade decoupled decoding and self-adaptive pseudo-labeling (SPL)), where class queries as latnet objetcness boost U-Recall by 1.4%-3.7% and known mAP by 1.1%-2.9% across Tasks 1-4 on S-OWODB (even without SPL, still achieving + 0.7%-2.1% in U-Recall).

C.3. Comparison with the PROB Baseline

PROB treats all unmatched queries as potential unknown objects by assigning pseudo-labels of 1. During inference, it leverages objectness scores to correlate with classification. In Tab. 2 of the main paper, the variant OWOBJ−S_{obj} assigns pseudo-labels to unknowns by sampling from a uniform distribution $\mathcal{U}(0, 1)$. To enable a more comprehensive and fair comparison with PROB, we introduce OWOBJ−$S_{obj(\mathbf{1})}$, which, similar to PROB, assigns an objectness score of 1 to all unmatched queries $\bar{\mathcal{Q}}$. We then apply the same change to OWOBJ−$\mathcal{L}_{obj}$, resulting in OWOBJ−$\mathcal{L}_{obj(\mathbf{1})}$, and compare both with PROB w/ or w/o correlation. Tab. 8 shows U-Recall/Known mAP (Task 1, M-OWODB). With correlation, both OWOBJ−$\mathcal{L}_{obj(\mathbf{1})}$ and PROB improve known mAP but lower U-Recall, while OWOBJ−$S_{obj(\mathbf{1})}$ shows a smaller increase in U-Recall, indicating that $\mathcal{L}_{obj}$ and $\mathcal{L}_{KL}$ together strengthen open-world objectness modeling.

D. Visualization

D.1. Open-World Object Detection (OWOD)

Fig. 6 presents qualitative results from COCO dataset, comparing how OW-DETR, PROB, and PROB+OWOBJ perform in detecting both known and unknown objects. Notably, OW-DETR demonstrates limitations in detecting unknown objects,

Table 6. State-of-the-art comparison for incremental object detection (iOD) on PASCAL VOC. We evaluate three settings using per-class AP and overall mAP@50 as metrics. The 10, 5 and 1 classes in gray background are introduced to a detector trained on the remaining 10, 15 and 19 classes, respectively. PROB+OWOBJ achieves favorable performance compared to existing OWOD approaches in all three settings.

10 + 10 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ILOD [41]	69.9	70.4	69.4	54.3	48	68.7	78.9	68.4	45.5	58.1	59.7	72.7	73.5	73.2	66.3	29.5	63.4	61.6	69.3	62.2	63.2
Faster ILOD [36]	72.8	75.7	71.2	60.5	61.7	70.4	83.3	76.6	53.1	72.3	36.7	70.9	66.8	67.6	66.1	24.7	63.1	48.1	57.1	43.6	62.1
ORE − (CC + EBUI) [20]	53.3	69.2	62.4	51.8	52.9	73.6	83.7	71.7	42.8	66.8	46.8	59.9	65.5	66.1	68.6	29.8	55.1	51.6	65.3	51.5	59.4
ORE − EBUI [20]	63.5	70.9	58.9	42.9	34.1	76.2	80.7	76.3	34.1	66.1	56.1	70.4	80.2	72.3	81.8	42.7	71.6	68.1	77	67.7	64.5
OW-DETR[17]	61.8	69.1	67.8	45.8	47.3	78.3	78.4	78.6	36.2	71.5	57.5	75.3	76.2	77.4	79.5	40.1	66.8	66.3	75.6	64.1	65.7
PROB [71]	70.4	75.4	67.3	48.1	55.9	73.5	78.5	75.4	42.8	72.2	64.2	73.8	76.0	74.8	75.3	40.2	66.2	73.3	64.4	64.0	66.5
PROB+OWOBJ	75.9	80.7	73.3	52.1	58.8	77.7	is 81.9	79.1	47.9	77.8	70.4	78.9	80.3	79.3	80.0	45.1	70.2	78.4	68.5	68.4	**69.9**
15 + 5 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ILOD [41]	70.5	79.2	68.8	59.1	53.2	75.4	79.4	78.8	46.6	59.4	59	75.8	71.8	78.6	69.6	33.7	61.5	63.1	71.7	62.2	65.8
Faster ILOD [36]	66.5	78.1	71.8	54.6	61.4	68.4	82.6	82.7	52.1	74.3	63.1	78.6	80.5	78.4	80.4	36.7	61.7	59.3	67.9	59.1	67.9
ORE − (CC + EBUI) [20]	65.1	74.6	57.9	39.5	36.7	75.1	80	73.3	37.1	69.8	48.8	69	77.5	72.8	76.5	34.4	62.6	56.5	80.3	65.7	62.6
ORE − EBUI [20]	75.4	81	67.1	51.9	55.7	77.2	85.6	81.7	46.1	76.2	55.4	76.7	86.2	78.5	82.1	32.8	63.6	54.7	77.7	64.6	68.5
OW-DETR [17]	77.1	76.5	69.2	51.3	61.3	79.8	84.2	81.0	49.7	79.6	58.1	79.0	83.1	67.8	85.4	33.2	65.1	62.0	73.9	65.0	69.4
PROB [71]	77.9	77.0	77.5	56.7	63.9	75.0	85.5	82.3	50.0	78.5	63.1	75.8	80.0	78.3	77.2	38.4	69.8	57.1	73.7	64.9	70.1
PROB+OWOBJ	82.8	80.0	82.4	60.1	68.0	79.9	90.0	86.4	54.4	83.1	64.2	77.3	85.1	80.3	80.1	42.1	73.2	61.8	77.9	68.7	**73.3**
19 + 1 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ILOD [41]	69.4	79.3	69.5	57.4	45.4	78.4	79.1	80.5	45.7	76.3	64.8	77.2	80.8	77.5	70.1	42.3	67.5	64.4	76.7	62.7	68.2
Faster ILOD [36]	64.2	74.7	73.2	55.5	53.7	70.8	82.9	82.6	51.6	79.7	58.7	78.8	81.8	75.3	77.4	43.1	73.8	61.7	69.8	61.1	68.5
ORE − (CC + EBUI) [20]	60.7	78.6	61.8	45	43.2	75.1	82.5	75.5	42.4	75.1	56.7	72.9	80.8	75.4	77.7	37.8	72.3	64.5	70.7	49.9	64.9
ORE − EBUI [20]	67.3	76.8	60	48.4	58.8	81.1	86.5	75.8	41.5	79.6	54.6	72.8	85.9	81.7	82.4	44.8	75.8	68.2	75.7	60.1	68.8
OW-DETR [17]	70.5	77.2	73.8	54.0	55.6	79.0	80.8	80.6	43.2	80.4	53.5	77.5	89.5	82.0	74.7	43.3	71.9	66.6	79.4	62.0	70.2
PROB [71]	80.3	78.9	77.6	59.7	63.7	75.2	86.0	83.9	53.7	82.8	66.5	82.7	80.6	83.8	77.9	48.9	74.5	69.9	77.6	48.5	72.6
PROB+OWOBJ	86.1	83.9	83.4	62.9	65.9	79.9	90.6	87.3	56.9	86.5	70.3	85.9	84.7	86.9	81.6	51.9	78.8	73.5	80.7	53.8	**75.8**

Table 7. The results of implementing our objectness modeling in both MEPU-FS [14] and CAT [33].

Task IDs (→)	**Task 1**		**Task 2**		**Task 3**		**Task 4**
	U-Recall ↑	Known mAP ↑	U-Recall ↑	Known mAP ↑	U-Recall ↑	Known mAP ↑	Known mAP ↑
MEPU-FS	37.9	74.3	35.8	54.3	35.7	46.2	41.2
MEPU-FS+OWOBJ	**39.7**	**77.4**	**37.8**	**57.2**	**38.1**	**49.0**	**43.5**
CAT (w/ SPL)	24.0	74.2	23.0	50.7	24.6	45.0	42.8
CAT (w/o SPL) +OWOBJ	26.1	75.4	23.7	51.5	26.0	45.6	43.2
CAT (w/ SPL) +OWOBJ	**27.7**	**77.1**	**24.9**	**53.3**	**27.0**	**49.0**	**43.9**

Table 8. Comparison with the PROB on Task 1, M-OWODB.

Method (→)	**PROB**		**OWOBJ**−$S_{obj(1)}$		**OWOBJ**−$\mathcal{L}_{obj(1)}$	
	U-Recall ↑	Known mAP ↑	U-Recall ↑	Known mAP ↑	U-Recall ↑	Known mAP ↑
w/ corr.	19.4	59.5	21.5	60.0	20.2	58.2
w/o corr.	21.1	39.3	21.0	43.9	22.3	41.3

primarily due to its reliance on unreliable heuristic methods in selecting unknown labels during training (*e.g.*, using the top 5 activation values from backbone features), leading to a low recall for unknown objects. While PROB avoids heuristic-based pseudo-labeling, it fails to effectively capture the objectness posterior distribution. As a result, the classifier lacks the ability to perceive the objectness of the observed data, frequently misclassifying most regions as unknown objects. This leads to a high recall for unknown objects, but it significantly compromises the performance of detecting known object. Our method jointly measures both the class and objectness distributions, effectively balancing high detection performance for both unknown and known objects.

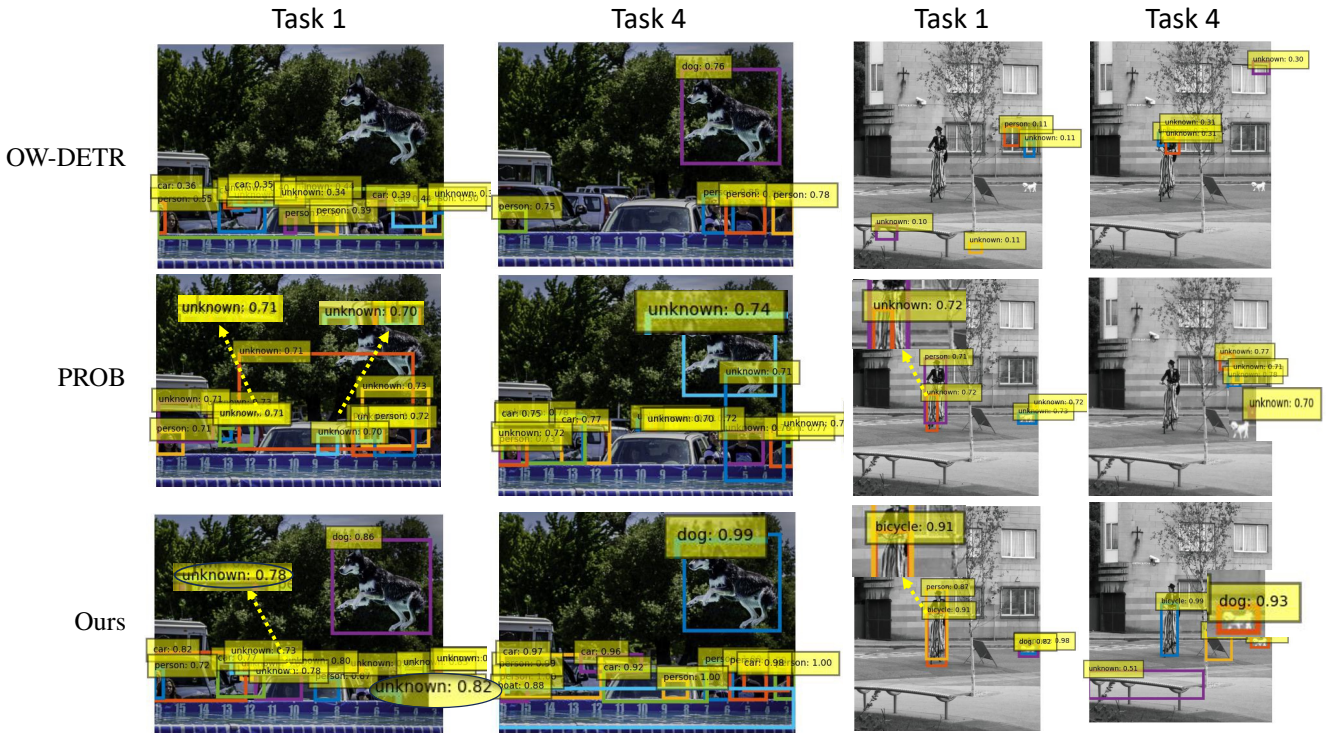


Figure 6. Qualitative examples on example images from the COCO test set on the first and last tasks. OW-DETR tends to miss unknown objects, leading to low unknown recall. RROB detects more unknown objects while suppressing the known detections. Our method achieves high detection performance for both unknown and known objects.

D.2. Few-Shot Object Detection (FSOD)

Fig. 7 illustrates the qualitative results of our approach on FSOD under the 5-shot protocol, showcasing detections for both base (top panel) and novel (bottom panel) classes. For base classes (*e.g.*, cup, carrot, vase, and oven), DeFRCN+OWOBJ significantly reduces redundant detections, improving the overall precision and efficiency of the detector (the second row). For novel classes, our method effectively recognizes unseen objects that were not part of the base training set (the last row). Notably, DeFRCN+OWOBJ not only improves the detection of novel objects but also achieves more accurate localization, as evidenced by tighter and more precise bounding boxes. These results demonstrate OWOBJ's strong generalization to novel object classes with minimal training samples, aligning with the overarching goal of few-shot object detection.

D.3. zero-shot Open-Vocabulary Object Detection (OVOD)

Fig. 8 provides a comparative visualization of box predictions generated by the baseline CORA (first and third rows) and CORA+OWOBJ (second and fourth rows) on the OV-LVIS dataset. The results clearly demonstrate the advantages of integrating OWOBJ into the baseline. CORA+OWOBJ successfully detects novel objects missed by the baseline. For example, in the last image of the second panel, CORA+OWOBJ successfully detects a smaller-scale *ring* that the baseline fails to identify, highlighting the improved sensitivity and generalization of our method, even for challenging and subtle objects. Moreover, the integration of the energy-based margin loss enables CORA+OWOBJ to assign higher confidence scores to unknown objects, distinguishing them more effectively from the background. This improvement is evident in the magnified regions that focus on the predicted bounding boxes, where CORA+OWOBJ provides more accurate and well-calibrated detections than the baseline.

D.4. Failure cases

Fig. 9 presents an analysis of the failure modes of CORA+OWOBJon the OV-LVIS validation set. Unrecognized novel objects are marked with red bounding boxes, showing the challenges that our approach still faces. Despite the advancements of OWOBJ some challenging scenarios remain where the model struggles to generalize effectively, particularly with novel

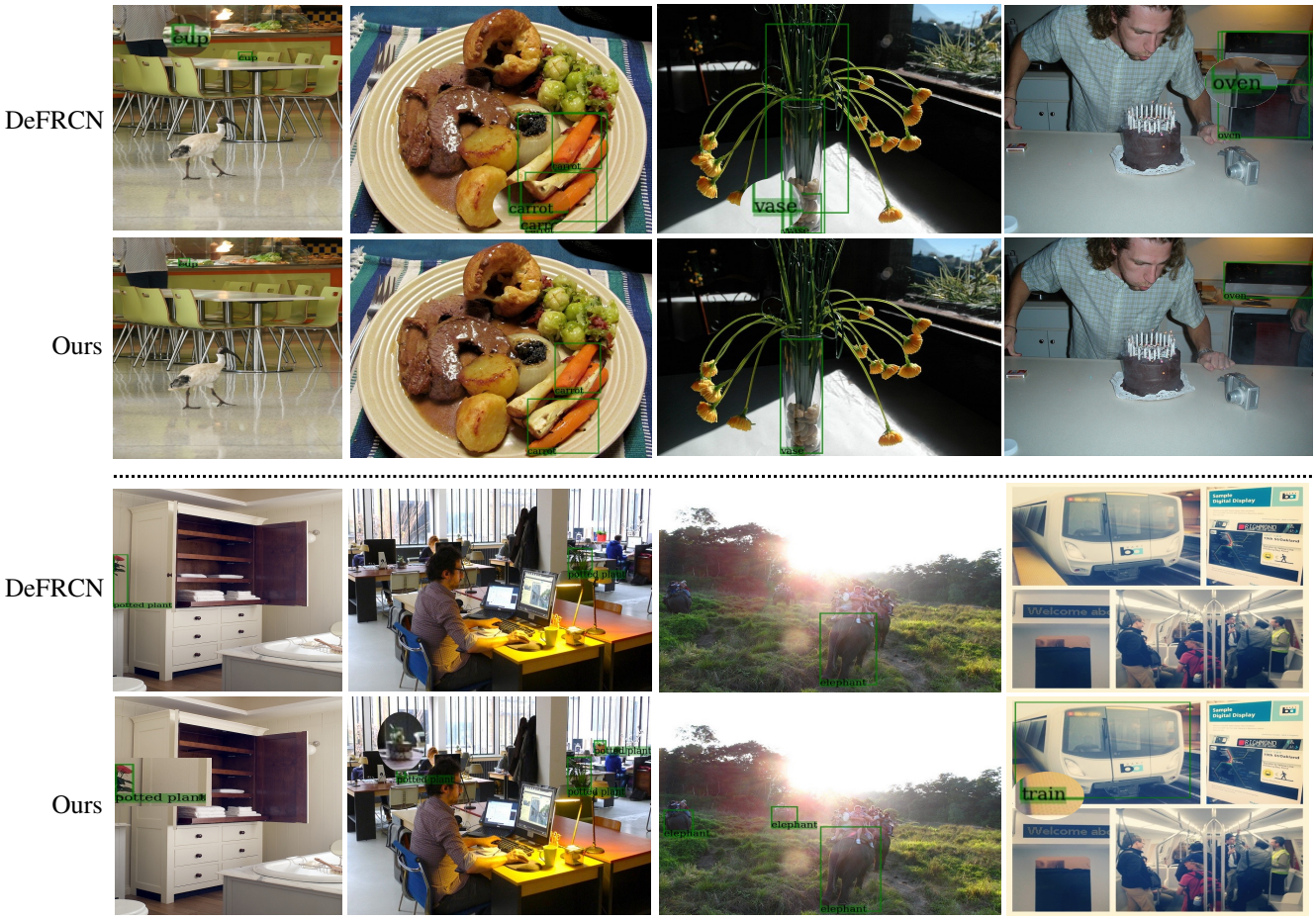


Figure 7. Visualization of FSOD detection results: Box predictions for base classes (top panel) and novel classes (bottom panel) under the 5-shot protocol. For base classes (*e.g.*, cup, carrot, vase, oven), ours (DeFRCN+OWOBJ) effectively reduces redundant detections (the second row). For novel classes, ours (DeFRCN+OWOBJ) improves the recognition of novel objects with more precise localization (the last row). Zoom in for the best view.

objects that exhibit low contrast, occlusion, or appear in cluttered environments.

E. Related Work

Open-World Object Detection (OWOD). OWOD is rendered more challenging than open set classification [3, 9, 18, 28] by the fact that instances of many unknown classes may be present but unlabeled in the training data. Since they are unlabeled, these unidentified instances are at risk of being learned as background [20]. The first OWOD method was ORE [20] which adapted the Faster-RCNN model by incorporating feature-space contrastive clustering, an RPN-based unknown detector, and an Energy-Based Unknown Identifier (EBUI). Yang *et al*. [53] introduced semantic topology to ensure the discriminative and consistent nature of feature representations. Zhao *et al*. [63] used Selective Search and a class-specific expelling function to compensate for over-confident activation boundaries. Transformer-based methods have shown significant promise in addressing OWOD. OW-DETR [17] was the first model to exhibit a deformable DETR (D-DETR)-serial detector [69] tailored for OWOD tasks. OW-DETR utilized a pseudo-labeling scheme to supervise unknown object detection, where unmatched object proposals with high backbone activations are selected as unknown objects (Top5 by default). This unknown object discovery strategy often selects parts of known proposals, or genuine background regions, however, leading to unreliable pseudo-labels for unknown objects, thereby resulting in low unknown object detection performance. Conversely, PROB [71]

Figure 8. Qualitative visualization of OV-LVIS validation set for novel classes. Rows 1 and 3 depict results generated by the baseline CORA, while Rows 2 and 4 display detection results produced by ours (CORA+OWOBJ). Zoom in for the best view.

does not identify unknown objects but instead trained a probabilistic objectness head alongside D-DETR. PROB fails to measure the objectness posterior, however, and this hinders the classifier's ability to differentiate real objects from non-objects. Without calibrating classification confidence with objectness scores during inference, OWOD performance degrades significantly. Moreover, PROB is limited to OWOD, and its extension to other novel object detection tasks remains unclear.

Few-shot Object Detection (FSOD). FSOD can be categorized into transfer learning and meta-learning paradigms. Meta-learning-based detectors employ a stage-wise, periodic meta-training paradigm to train a meta-learner, facilitating knowledge transfer from base classes. Meta R-CNN [52] re-weights query RoI features by support prototypes in the detection head. FSOD-ARPN [13] uses a channel-wise attention for RPN and multi-relation detector. PNSD [59] extends FSOD-ARPN by contracted autocorrelation matrix to improve attention. KFSOD [61] and TENET [60] extend PNSD with kernel/tensor representations. With the balanced dataset introduced in TFA [47], fine-tuning-based detectors are outperforming meta-learning-based methods. MPSR [48] deals with scale invariance by ensuring the detector is trained over multiple scales of positive samples. NP-RepMet [54] introduces a negative- and positive-representative learning framework via triplet losses that bootstrap the classifier. Similarly, FSCE [43] aims to decrease instance similarity between objects belonging to different categories by adding a secondary branch to the primary RoI head, which is trained via supervised contrastive learning. De-FRCN [37] proposes to perform stop-gradient between the RPN and the backbone to deal with the inconsistent optimization goals between them.

Zero-Shot/Open-Vocabulary Object Detection (OVOD). Scaling up box-level annotations is costly and labor-intensive; however, zero-shot detection facilitates identifying novel categories absent from the training data. OVR-CNN [58] first proposes the OVOD benchmark to bridge the performance gap between Zero-Shot and supervised learning. Specifically, it integrates image-language alignment knowledge from large pre-trained vision-language models (VLMs), such as CLIP [38], into object detectors. Current techniques are generally divided into four strategies: pseudo-labeling [58, 62, 64, 65], distillation [11, 15], conditional matching [49, 57] and parameter transfer [25, 49]. Although such methods leverage the

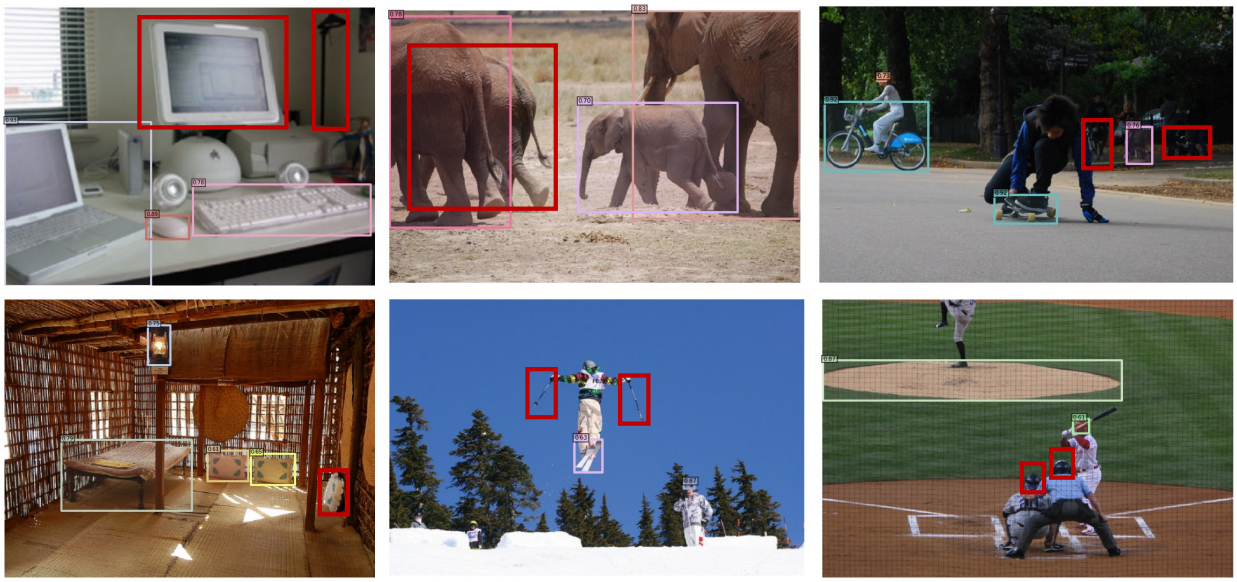


Figure 9. Analysis of failure modes of CORA+OWOBJ on the OV-LVIS validation set. Unrecognized novel objects are marked with red bounding boxes. Zoom in for the best view.

strong zero-shot recognition capabilities of VLMs to handle open vocabularies, open vocabulary detectors primarily rely on base detection data with box-level annotations during optimization, inherently biasing them toward base categories. As a result, novel objects are easily regarded as background or base categories, underscoring the need for general objectness.