OW-DAR: Dual-Granularity Adaptive Reconstruction-Error Modeling for Open-World Object Detection

Linhua Ye[1], **Xing Xi**[1], **Ronghua Luo**[1*]

[1]School of Computer Science and Engineering, South China University of Technology
cslinhuaye@mail.scut.edu.cn, xxyzll@yeah.net, rhluo@scut.edu.cn

Abstract

Open-world object detection (OWOD) aims to detect known and unknown objects in dynamic environments. However, only known classes are labeled during training, making it challenging for detectors to recognize unknown objects during inference. Existing methods typically rely on supervision from known categories, leading models to overconfidently misclassify visually similar unknowns as known, and dissimilar ones as background. This known-class prior bias limits the model's ability to detect unknown objects. In this paper, we propose a novel method, OW-DAR, which enhances foreground-background separability through collaborative fine-grained and coarse-grained modeling. At the fine-grained level, we propose Fine-grained Masked Reconstruction (FMR), which randomly masks regions of the feature map to guide the reconstruction toward semantic structures, rather than memorizing low-level patterns. At the coarse-grained level, we propose Adaptive Region-based Error Aggregation (AREA), which operates on object proposals to aggregate reconstruction errors. This enables the model to attend to semantically ambiguous foreground-background boundaries while suppressing the influence of local outliers during optimization. Finally, we leverage robust reconstruction errors to perform unsupervised foreground-background modeling, enabling probabilistic estimation for potential unknown objects. We validate the effectiveness of OW-DAR on standard OWOD benchmark. Experimental results demonstrate that OW-DAR consistently outperforms existing state-of-the-art methods, achieving a +18.8 improvement in unknown object recall (U-Recall).

Code — https://github.com/lllhye/OW-DAR

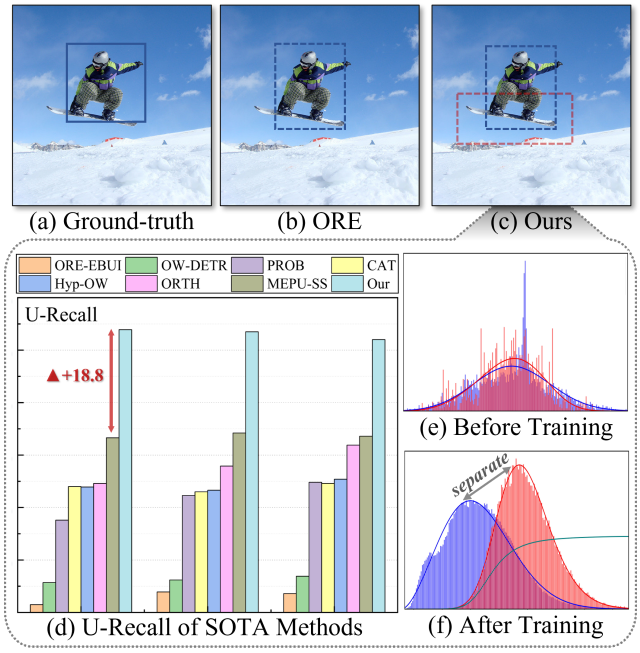


Figure 1: (a) Ground-truth annotations. (b)-(c) Predictions of unknown (red) and known (blue) objects from ORE and our method. (d) U-Recall on the standard OWOD benchmark for SOTA methods. (e) Reconstruction error distribution before training, where foreground and background are not clearly distinguishable. (f) Our method enables better separation between foreground and background after training.

Introduction

Object detection is a fundamental task in computer vision, widely applied in fields such as autonomous driving (Ma et al. 2022; Li et al. 2022), robotics (Nie et al. 2023; Bai et al. 2024), and medical imaging (Elakkiya et al. 2022; Karri et al. 2022). However, most existing object detection methods operate under the closed-world assumption, where models are trained on a static and predefined set of object categories, a condition that rarely holds in real-world environments. To improve the adaptability of object detectors to unknown categories, Continual Object Detection (COD) (Menezes et al. 2023) has been proposed, enabling models to incrementally expand their category space. However, training under COD settings often involves shifts in the input distribution and is prone to catastrophic forgetting (Kirkpatrick et al. 2017; Doan et al. 2021), making it difficult for models to retain knowledge of previously learned categories.

While COD allows for incremental learning of known categories, it is not designed to handle entirely unknown classes. To this end, Open World Object Detection (OWOD) (Joseph et al. 2021) introduces a more realistic setting, where models are trained with labeled known categories

*Corresponding author: Ronghua Luo.

while being required to detect and incrementally learn unknown categories without any supervision. Despite its stronger alignment with real-world demands, OWOD introduces new challenges. In particular, due to the absence of supervision for unknown objects, models often struggle to recognize such instances effectively and may erroneously classify them as background.

To address this issue, existing OWOD methods commonly rely on pseudo-labeling mechanisms to facilitate the discovery of unknown objects. For instance, ORE (Joseph et al. 2021) selects background regions with high objectness scores as potential unknowns, while OW-DETR (Gupta et al. 2022) leverages attention mechanisms to supervise high-confidence candidate regions. However, these approaches heavily depend on objectness knowledge learned from known categories, which exacerbates the problem of known-class prior bias—making it difficult f or models to distinguish unknown objects from background regions. This bias manifests in two ways: when an unknown object is significantly different from known categories, it is often ignored as background; conversely, when it appears similar in the feature space, it is likely to be misclassified as a known category. As illustrated in Fig. 1(b), although the skateboard (an unknown object) is clearly visible, ORE (Joseph et al. 2021) fails to detect it correctly due to its low similarity to any known category (e.g., "person") in the feature space.

To overcome this limitation, we draw inspiration from human perception theories, particularly predictive coding (Clark 2013; Friston 2010) and global-local processing (Navon 1977), which suggest that humans detect unknown stimuli by minimizing prediction errors across multiple levels of granularity. At a coarse (global) scale, overall structure and saliency are perceived, while at a fine (local) scale, detailed differences are discerned. This dual-granularity perception facilitates robust unknown object recognition under complexity and ambiguity.

Motivated by this mechanism, we propose a novel OWOD method, OW-DAR, explicitly modeling foreground-background differences using dual-granularity reconstruction error modeling. As shown in Fig. 1(c-f), this design facilitates more accurate and robust estimation of unknown object probabilities. Our method significantly improves U-Recall by +18.8 percentage points over prior state-of-the-art (SOTA) methods (Joseph et al. 2021; Gupta et al. 2022; Zohar, Wang, and Yeung 2023; Ma et al. 2023a; Doan et al. 2024; Sun, Li, and Mu 2024; Fang et al. 2025), while employing a purely convolutional architecture without introducing additional parameters, thereby offering a significant advantage in inference speed.

Our main contributions are as follows:

- We enhance unknown object detection through collaborative modeling at fine-grained and coarse-grained levels, and enable unbiased probabilistic estimation for them.
- We propose the Fine-grained Masked Reconstruction (FMR), which performs element-wise masking to selectively obscure feature inputs and guides the model to focus on reconstructing semantically structures.
- We propose the Adaptive Region-based Error Aggregation (AREA), which adaptively aggregates reconstruction errors over object proposals, enabling the model to distinguish ambiguous foreground-background boundaries and suppress noise from local outliers.
- Extensive experiments on standard OWOD benchmarks demonstrate OW-DAR consistently outperforms existing SOTA methods, achieving 52.1 U-Recall and 75.1 mAP.

Related Works

Open-World Object Detection

Object detection has achieved remarkable progress with model families such as R-CNN (Girshick 2015; Girshick et al. 2014; He et al. 2017; Ren et al. 2016), YOLO (Redmon 2016; Ge et al. 2021; Khanam and Hussain 2024), and DETR (Carion et al. 2020a,b; Zhao et al. 2024). However, their ability to detect unknown objects in complex real-world scenarios remains limited. To address this, Open World Object Detection (OWOD) is introduced by ORE (Joseph et al. 2021), posing a challenging task in which the detector must recognize unknown objects in fully open environments without prior annotations and support incremental learning for new categories. Most OWOD methods rely on heuristic cues to localize unknowns. ORE selects high-objectness proposals from the RPN; OW-DETR (Gupta et al. 2022) assigns pseudo-labels to top-k regions ranked by average scores; CAT (Ma et al. 2023a) combines cascaded decoding with adaptive pseudo-labeling to reduce category bias. PROB (Zohar, Wang, and Yeung 2023) models known instances to discover unknowns; Hyp-OW (Doan et al. 2024) adopts similarity-based relabeling; ORTH (Sun, Li, and Mu 2024) enforces orthogonality in feature space to reduce object-class entanglement. Although these approaches have made notable progress in OWOD, they all inherently depend on supervision from known categories, making it difficult to avoid the known-class prior bias. These limitations motivate us to explore unsupervised foreground-background modeling based on fine-grained and coarse-grained collaboration, enabling unbiased probabilistic estimation for potential unknown objects in open-world scenarios.

Reconstruction-based OOD Detection

Out-of-Distribution (OOD) detection refers to identify and reject test samples that fall outside the distribution of a model's training data. Reconstruction-based methods (Denouden et al. 2018; Zhou 2022; Jiang et al. 2023) assume that an encoder-decoder architecture trained exclusively with in-distribution (ID) data typically yields higher reconstruction errors when encountering OOD samples. Consequently, ID and OOD samples can be effectively discriminated by analyzing reconstruction errors during inference. Denouden et al. (Denouden et al. 2018) introduced the Mahalanobis distance in the latent space to more accurately capture OOD samples that are distant from the ID distribution yet remain near the manifold of the latent representation. Zhou et al. (Zhou 2022) further formalized reconstruction-based OOD detection into a quadruplet domain translation framework, significantly enhancing detection performance

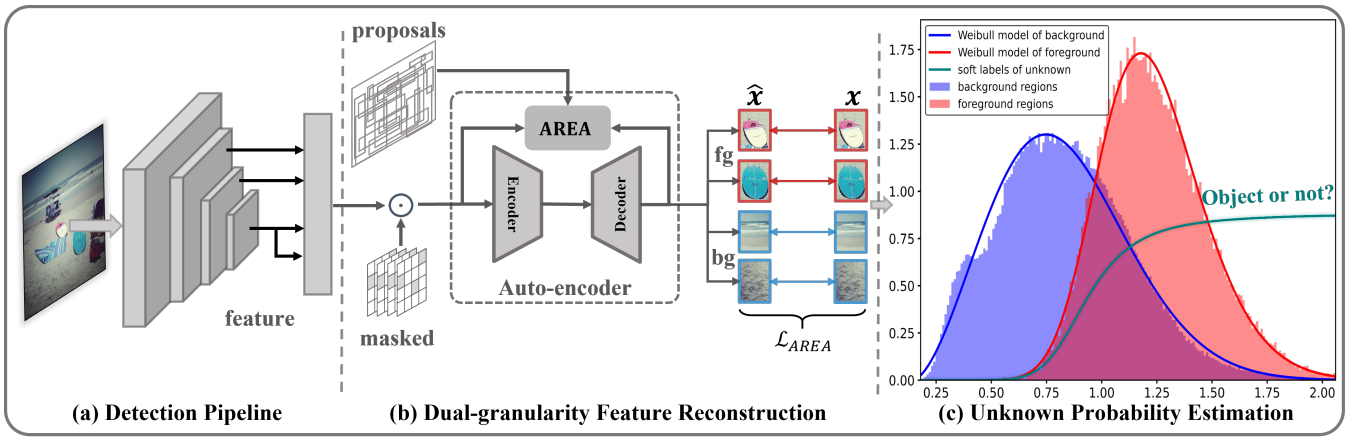


Figure 2: Overall architecture of OW-DAR. (a) We build on a Faster R-CNN architecture with FPN to extract multi-scale features for detecting both known and unknown objects. (b) A dual-granularity auto-encoder reconstructs element-wise masked feature maps at a fine-grained level and aggregates reconstruction errors at the object proposal level via AREA to enhance foreground-background separability. (c) Robust reconstruction errors are employed to perform unsupervised modeling of foreground and background, enabling probabilistic estimation of potential unknown objects from noisy proposals.

by employing semantic reconstruction, data certainty decomposition, and normalized L2 distance strategies. Additionally, Jiang et al. (Jiang et al. 2023) proposed READ, which maps pixel-level reconstruction errors into the latent feature space of a classifier, effectively integrating the strengths of auto-encoders and classification models. Inspired by these reconstruction-based OOD detection approaches, we revisit the potential application of reconstruction error modeling in OWOD. Unlike traditional image-level reconstruction paradigms, our work focuses specifically on the discriminative capability of reconstruction errors to differentiate foreground and background regions.

Methodology

Given the inherent differences in visual structure and occurrence frequency between foreground and background, we leverage reconstruction error as a discriminative signal for unsupervised modeling. Background regions typically produce low errors, while structurally diverse foregrounds yield higher errors (Fang et al. 2025). Based on this insight, we build OW-DAR on the Faster R-CNN (Ren et al. 2016) framework with a Feature Pyramid Network (FPN) (Lin et al. 2017), and design a dual-granularity encoder-decoder architecture to enhance foreground-background separability. As shown in Fig. 2, the model leverages reconstruction error across fine and coarse levels to support unsupervised probabilistic modeling of potential unknown objects.

Fine-grained Masked Reconstruction

Traditional auto-encoders aim to reconstruct input data in an unsupervised manner, with latent spaces primarily capturing global statistical patterns rather than localized semantic structures (Berthelot et al. 2018; Hinton and Salakhutdinov 2006). Low-complexity regions such as background regions are easily reconstructed using simple patterns, while foreground objects with fine semantic details are often smoothed due to compression in the low-dimensional latent space, leading to limited reconstruction error differences between foreground and background. A natural solution might be to employ a masked reconstruction strategy, such as the masked auto-encoder (MAE) (He et al. 2022), which uses random masking on input features to compel the model to learn global context and predict missing regions. However, the global random masking strategy in MAE does not adequately focus on distinguishing between foreground and background. Its objective is to learn holistic image-level representations rather than enhance the separability of reconstructed features across foreground and background.

To address this limitation, we propose Fine-grained Masked Reconstruction (FMR), which performs element-wise masking in the feature space with added noise perturbation. This guides the model to focus on local semantic context rather than memorizing low-level patterns. This strategy maintains low reconstruction errors in structurally regular background regions while amplifying errors in foreground regions with complex semantics and frequent occlusions. The resulting contrast enhances foreground-background separability and improves the recall of unknown objects.

We employ the FPN to extract multi-scale feature maps from input image $\mathbf{I} \in \mathbf{R}^{H_0 \times W_0 \times 3}$. For simplicity, we describe the masking process at a single representative scale, where the feature map is denoted as $\mathbf{X} \in \mathbf{R}^{H_F \times W_F \times C}$. We perform element-wise random masking and generate a binary mask $\mathbf{M} \in \{0,1\}^{H_F \times W_F}$ aligned with $\mathbf{X}$, where each element is sampled as:

$$\mathbf{M}_{i,j} = \begin{cases} 0, & \text{with probability } r, \\ 1, & \text{otherwise}, \end{cases} \tag{1}$$

where r is the element-wise masking ratio applied in FMR. We apply the binary mask $\mathbf{M}$ to the feature map $\mathbf{X}$ via element-wise multiplication to obtain the masked feature

map $\mathbf{X}_{mask}$:

$$\mathbf{X}_{mask} = \mathbf{X} \odot \mathbf{M}, \tag{2}$$

where $\odot$ denotes element-wise multiplication. The masked features $\mathbf{X}_{\text{mask}}$ are subsequently processed by an autoencoder, comprising an encoder $\mathcal{E}(\cdot)$ and a decoder $\mathcal{D}(\cdot)$, to reconstruct the original unmasked features:

$$\hat{\mathbf{X}} = \mathcal{D}\left(\mathcal{E}\left(\mathbf{X}_{mask}\right)\right), \tag{3}$$

where, $\hat{\mathbf{X}} \in \mathbf{R}^{H_F \times W_F \times C}$ denotes the reconstructed feature map from the masked input. Finally, we calculate the element-wise reconstruction error $\mathbf{E}$ as follows:

$$\mathbf{E} = \|\hat{\mathbf{X}} - \mathbf{X}\|, \quad \mathbf{E} \in \mathbf{R}^{H_F \times W_F \times 1}. \tag{4}$$

The error captures local reconstruction difficulty, with foreground regions producing higher values due to structural complexity and background regions producing lower values due to their simplicity and repetition.

Adaptive Region-based Error Aggregation

Although fine-grained reconstruction enhances the discriminability between foreground and background, directly relying on element-wise reconstruction errors to detect unknown objects may lead to misclassification in complex backgrounds. Specifically, some background regions with repetitive textures, local occlusions, or structural variations may still produce high reconstruction errors, even though they do not correspond to real foreground objects, which can lead to false detections. Furthermore, extremely high values at individual pixels can dominate the loss calculation, leading to unstable training and reduced model robustness.

To address these challenges, we propose Adaptive Region-based Error Aggregation (AREA). This method aggregates reconstruction errors at the object proposal level and dynamically adjusts their contributions based on region-level aggregation, enhancing the model's ability to distinguish between foreground and background in ambiguous regions. Unlike existing methods that treat all element-wise errors equally, AREA explicitly models the structured error patterns within candidate regions, thereby improving the recognition accuracy of potential unknown objects.

We adopt Selective Search (Uijlings et al. 2013) as the default unsupervised proposal generator. Notably, the strength of OW-DAR lies not in the proposal strategy itself, but in its ability to discriminate between foreground and background regions, making it robust across diverse proposal sources. Given each input image, we extract N candidate regions, denoted as $\mathcal{B} = \{\mathbf{b}_i\}_{i=1}^{N}$. Each candidate box $\mathbf{b}_i$ is defined by its top-left coordinate (x_i, y_i) and its width and height (w_i, h_i). We then use ROIAlign to extract region-level reconstruction errors $\mathbf{E}_{\mathbf{b}_i}$ from the element-wise feature reconstruction error map $\mathbf{E}$:

$$\mathbf{E}_{\mathbf{b}_i} = \text{ROIAlign}\left(\mathbf{E}, \mathbf{b}_i\right). \tag{5}$$

Subsequently, we compute the mean of pixel-wise errors within each region to suppress noise and capture region-level semantic uncertainty:

$$\widetilde{e}_i = \frac{1}{|b_i|} \sum_{j \in b_i} e_j. \tag{6}$$

This aggregation process suppresses fine-grained fluctuations in background reconstruction errors, thus allowing the model to attend to semantically informative region-level error patterns rather than unstable variations at individual spatial locations. The corresponding loss function is defined as follows:

$$\mathcal{L}_{AREA} = \frac{1}{N} \sum_{i=1}^{N} \exp\left(-\alpha \widetilde{e}_i\right) \cdot \widetilde{e}_i^2, \tag{7}$$

where α controls the degree of dynamic adjustment, allowing the model to suppress the influence of extreme reconstruction outliers and progressively focus on foreground-background boundary regions during training, rather than being dominated by extreme reconstruction errors.

Unknown Probability Estimation

We observe that the reconstruction errors of foreground and background exhibit a skewed distribution, indicating asymmetry in their error characteristics. To model this, we adopt asymmetric Weibull distributions for foreground and background (denoted as WBfg and WBbg), defined as follows:

$$\text{WB}(E; \lambda, k) = \frac{k}{\lambda} \left(\frac{E}{\lambda}\right)^{k-1} e^{-\left(\frac{E}{\lambda}\right)^k}. \tag{8}$$

After modeling the foreground and background distributions, we employ a soft-labeling mechanism to estimate the likelihood that a potential unknown region corresponds to a true unknown object. The soft label is computed as:

$$s(E_{\text{uk}}) = \left(\frac{\text{WB}_{\text{fg}}(E_{\text{uk}})}{\text{WB}_{\text{bg}}(E_{\text{uk}}) + \text{WB}_{\text{fg}}(E_{\text{uk}})}\right)^{\gamma}, \tag{9}$$

where $E_{\text{uk}} \in \mathbf{R}^{H_F \times W_F \times 1}$ denotes the reconstruction error map of a candidate unknown object. $\text{WB}_{\text{fg}}(\cdot)$ and $\text{WB}_{\text{bg}}(\cdot)$ represent the probability density functions (PDFs) of the asymmetric Weibull distributions for the foreground and background regions, respectively. The hyperparameter γ controls the sharpness of the soft label distribution: as $\gamma \to \infty$, all pseudo labels are suppressed; conversely, as $\gamma \to 0$, all pseudo labels are treated as true unknowns.

Experiments

Details of the datasets, evaluation metrics, implementation, and additional experimental analyses are provided in the Appendix.

Comparison With State-of-the-art Methods

We compare our proposed OW-DAR with recent SOTA methods (Joseph et al. 2021; Gupta et al. 2022; Ma et al. 2023a; Zohar, Wang, and Yeung 2023; Doan et al. 2024; Sun, Li, and Mu 2024; Fang et al. 2025; He et al. 2024), with results summarized in Tab. 1. ORE-EUBI (Joseph et al. 2021) removes the energy model from ORE due to data leakage. MEPU-SS is a MEPU (Fang et al. 2025) variant without the FreeSOLO (Wang et al. 2022) foundation model, and PROB+SAM denotes PROB (Zohar, Wang, and Yeung 2023) with the SAM (Kirillov et al. 2023) module, as reported in the SGROD (He et al. 2024) paper. As pointed

Task IDs (→)	Task 1		Task 2				Task 3				Task 4		
	U-Recall	mAP (↑)	U-Recall	mAP (↑)			U-Recall	mAP (↑)			mAP (↑)		
	(↑)	Current known	(↑)	Previously known	Current known	Both	(↑)	Previously known	Current known	Both	Previously known	Current known	Both
ORE-EBUI (CVPR2021)	1.5	71.4	3.9	61.0	30.9	45.6	3.6	43.1	32.2	39.5	33.6	26.3	31.8
OW-DETR (CVPR2022)	5.7	73.1	6.2	65.0	29.0	46.0	6.9	46.7	25.7	39.7	38.2	28.1	33.1
CAT (CVPR2023)	24.0	74.2	23.0	67.6	35.5	50.7	24.6	51.2	32.6	45.0	45.4	35.1	42.8
PROB (CVPR2023)	17.6	73.5	22.3	66.3	36.0	50.4	24.8	47.8	30.4	42.0	42.6	31.7	39.9
Hyp-OW (AAAI2024)	23.9	72.7	23.3	-	-	50.6	25.4	-	-	46.2	-	-	44.8
ORTH (CVPR2024)	24.6	71.6	27.9	64.0	39.9	51.3	31.9	52.1	42.2	48.8	48.7	38.8	46.2
MEPU-SS (TNNLS2025)	33.3	74.2	34.2	67.5	41.0	53.6	33.6	50.0	37.5	45.8	43.2	33.5	40.8
Ours: OW-DAR	**52.1**	**75.1**	**51.6**	**69.8**	**43.5**	**55.8**	**50.2**	**53.1**	**42.9**	**49.7**	**48.9**	**39.7**	**46.6**
MEPU-FS (TNNLS2025)	37.9	74.3	35.8	68.0	41.9	54.3	35.7	50.2	38.3	46.2	43.7	33.7	41.2
PROB+SAM	46.8	71.4	48.0	63.6	29.3	45.6	45.6	44.8	26.8	38.8	38.5	26.2	35.4
SGROD (TIP2024)	48.0	73.2	48.9	64.7	36.7	50.0	47.7	47.4	32.2	42.4	42.5	32.6	40.0
Ours: OW-DAR+SAM	**54.0**	**74.8**	**53.2**	**69.2**	**42.3**	**54.8**	**51.6**	**52.3**	**42.3**	**48.9**	**48.7**	**39.1**	**46.3**

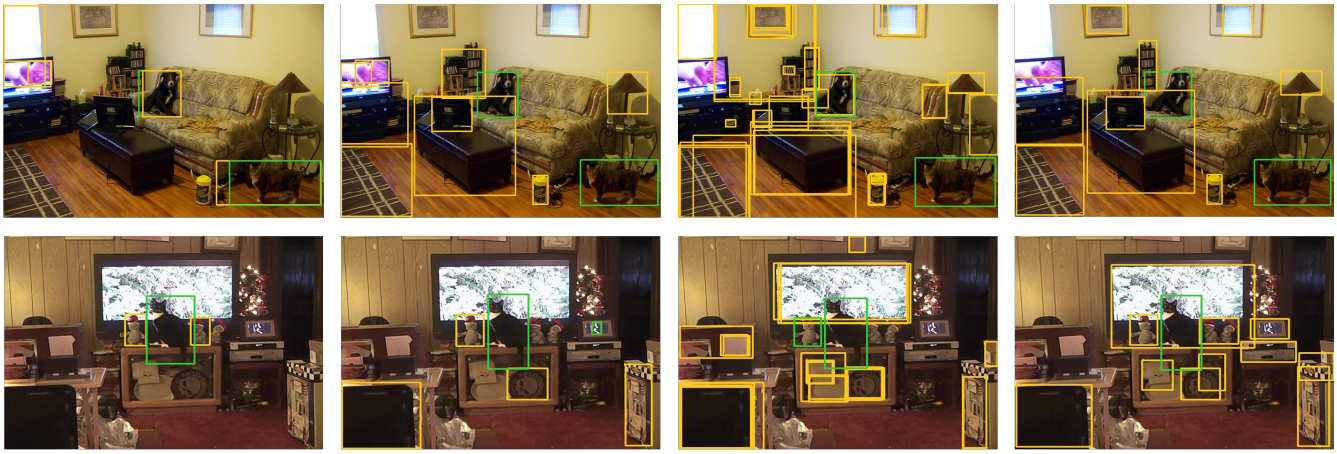
Table 1: Comparison of open-world object detection performance (U-Recall). The table compares our OW-DAR with previous SOTA methods. The upper half compares methods without foundation models, while the lower half compares methods with foundation models. U-Recall measures recall on unknown categories, and mAP evaluates detection on known categories. Current Known, Previously Known, and Both indicate newly introduced categories, categories from earlier tasks, and all categories encountered so far, respectively. OW-DAR shows significant improvements in recall for unknown categories and achieves competitive detection accuracy for known categories.

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 (CVPR2021)	1.5	0.0240	2486	3.9	0.0400	6608	3.6	0.0260	6896
OW-DETR (CVPR2022)	5.7	0.0290	12721	6.2	0.0410	14970	6.9	0.0240	9197
CAT (CVPR2023)	24.0	0.0230	2097	23.0	0.0400	5784	24.6	0.0210	3545
PROB (CVPR2023)	17.6	0.0206	2014	22.3	0.0309	3358	24.8	0.0176	1545
MEPU-SS (TNNLS2025)	33.3	0.0200	1753	34.2	0.0280	3352	33.6	0.0200	2883
Our:OW-DAR	**52.1**	**0.0186**	**1352**	**50.8**	**0.0242**	**2002**	**50.2**	**0.0152**	**1015**
MEPU-FS (TNNLS2025)	37.9	0.0200	1710	35.8	0.0270	3197	35.7	0.0200	2862
PROB+SAM	46.8	0.0341	6109	48.0	0.0499	6520	45.6	0.0176	1545
SGROD (TIP2024)	48.0	0.0266	2522	48.9	0.0329	2294	47.7	0.0201	1382
Our: OW-DAR+SAM	**54.0**	**0.0192**	**1480**	**53.2**	**0.0253**	**2132**	**51.6**	**0.0163**	**1128**

Table 2: Analysis of unknown object confusion on S-OWODB. The table compares methods using metrics including U-Recall, WI, and A-OSE. Our method achieves SOTA results in both U-Recall and A-OSE, while exhibiting competitive WI relative to other methods. Note that these metrics are omitted for Task 4, as all 80 classes become known at this stage.

Task IDs (→)	Task 1		Task 2				Task 3				Task 4		
	U-Recall	mAP (↑)	U-Recall	mAP (↑)			U-Recall	mAP (↑)			mAP (↑)		
	(↑)	Current known	(↑)	Previously known	Current known	Both	(↑)	Previously known	Current known	Both	Previously known	Current known	Both
Base Model	33.3	74.2	34.2	67.5	41.0	53.6	33.6	50.0	37.5	45.8	43.2	33.5	40.8
OW-DAR-FMR	45.6	74.8	43.9	68.9	42.6	55.2	42.6	52.7	41.8	49.1	47.5	39.1	45.4
OW-DAR-AREA	48.2	74.6	47.2	68.5	42.1	54.6	45.9	52.2	41.3	48.5	47.2	37.5	44.8
Final: OW-DAR	**52.1**	**75.1**	**51.6**	**69.8**	**43.5**	**55.8**	**50.2**	**53.1**	**42.9**	**49.7**	**48.9**	**39.7**	**46.6**

Table 3: Component ablation study. Comparison on S-OWODB based on mAP for known categories and U-Recall for unknown categories. OW-DAR-FMR is our model without the fine-grained FMR module. OW-DAR-AREA is our model without the coarse-grained AREA module.



ORTH FPS: 31.84 | MEPU FPS: 32.25 | SGROD FPS: 24.51 | **Ours** FPS: 35.42

Figure 3: Qualitative analysis. We compare representative SOTA methods including ORTH, MEPU-SS, and SGROD. Inference speed (FPS) is measured on a single RTX 3090 GPU. Green boxes denote detected known objects, while yellow boxes highlight predictions of potential unknown objects.

Method	U-Recall (↑)	K-mAP (↑)	WI (↓)	A-OSE (↓)
FreeSOLO	52.3	74.5	0.0194	1532
SAM	**54.0**	74.8	0.0192	1480
RandBox	51.8	**75.2**	**0.0185**	1368
Selective Search	52.1	75.1	0.0186	**1352**

Table 4: Comparison of different unsupervised region proposal methods on Task 1. Our method adopts Selective Search by default.

Method	U-Recall (↑)	K-mAP (↑)	WI (↓)	A-OSE (↓)
ℓ_1-norm	47.8	74.5	0.0198	1708
ℓ_2-norm	48.3	74.2	0.0196	1683
Huber	49.1	74.6	0.0197	1587
Ours	**52.1**	**75.1**	**0.0186**	**1352**

Table 5: Comparative analysis of different reconstruction loss functions in Task 1.

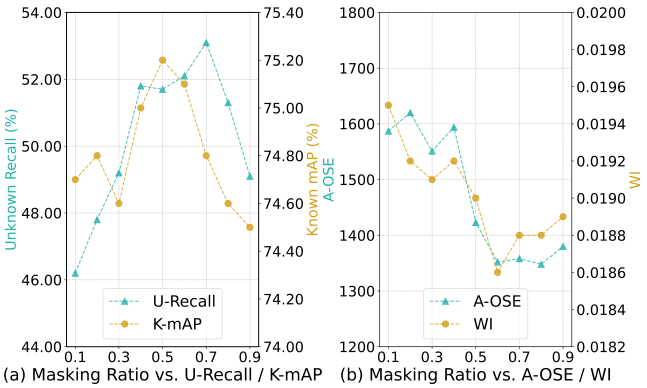


Figure 4: Sensitivity analysis on the masking ratio in FMR. (a) Impact on detection metrics (U-Recall and K-mAP). (b) Impact on confusion metrics (A-OSE and WI).

out by FOMO (Zohar et al. 2023), foundation models (e.g., CLIP (Radford et al. 2021) and SAM), when pretrained on large-scale web data, can cause serious data leakage when applied to OWOD datasets, especially for CLIP-based methods. Therefore, we exclude SKDF (Ma et al. 2023b), which uses CLIP as its backbone, from comparison.

It should be noted that although our extended experiments include SOTA methods based on SAM-generated pseudo labels, these methods still suffer from data leakage. Unlike CLIP, SAM does not encode explicit category-level semantic knowledge. Its data leakage primarily manifests at the masking level. Due to data leakage from foundation model-based pseudo labels, MEPU provides two variants: MEPU-FS (with FreeSOLO) and MEPU-SS (without it). Under the SAM-based setting, we compare with MEPU-FS and SGROD, which also leverage foundation model knowledge.

Identifying Potential Objects (T1) Tab. 1 compares OW-DAR with recent SOTA methods on the OWOD benchmark. For unknown-class recall, compared to ORTH (without foundation model), OW-DAR achieves a +27.5 gain in U-Recall. It also outperforms SGROD (with foundation model) by +6.0, and MEPU-SS (without FreeSOLO) by +18.8. For known-class detection, OW-DAR remains competitive, surpassing Hyp-OW and ORTH by +2.4 and +3.5 mAP, respectively. Notably, OW-DAR outperforms all these methods without relying on any vision foundation model. To assess the effect of foundation models, we introduce SAM-based supervision. With SAM, OW-DAR achieves a marginal U-Recall improvement of approximately +2.0. In contrast, PROB gains approximately +30.0 after incorporating SAM, indicating potential data leakage from pretrained foundation models. These results indicate that OW-DAR's improvements mainly arise from its dual-granularity reconstruction-error modeling and probabilistic estimation of unknown objects, rather than reliance on external founda-

tion model knowledge.

Incremental learning (T2-T4) As shown in Tab. 1, OW-DAR consistently outperforms ORTH and MEPU-SS across Tasks 2 to 4. It achieves superior U-Recall while maintaining strong detection performance on current known categories, demonstrating its ability to effectively balance unknown object discovery and known-class recognition.

As shown in Tab. 2, we further evaluate the confusion between known and unknown categories. OW-DAR not only achieves a significant improvement in U-Recall but also establishes new SOTA performance on WI (Dhamija et al. 2020) and A-OSE (Miller et al. 2018), indicating its ability to recall more unknown objects while effectively distinguishing between known and unknown categories.

Ablation Study

Component ablation study. (Tab. 3) Removing FMR leads to a clear drop in open-set performance (U-Recall), indicating its effectiveness in reducing misclassification of unknown objects as background or known classes. Similarly, removing AREA degrades both closed-set (mAP) and open-set (U-Recall) performance, further validating its role in mitigating foreground-background confusion and alleviating known-class prior bias.

Unsupervised Region Proposal (Tab. 4) We evaluate several representative unsupervised region proposal methods, including the foundation models FreeSOLO (Wang et al. 2022) and SAM (Kirillov et al. 2023), as well as non-foundation alternatives such as Selective Search (Uijlings et al. 2013) and RandBox (Wang et al. 2023). Among these, Selective Search serves as the default proposal generator in our method. Notably, although Selective Search is a well-established classical method and tends to generate many background regions, it still achieves a U-Recall of 52.1, validating the effectiveness of our dual-granularity reconstruction-error modeling. In addition, RandBox, which serves as a fully random proposal generator without prior knowledge, still achieves a competitive U-Recall of 51.8 in our framework. This result indicates that OW-DAR remains effective even without semantically guided proposals because it provides stronger foreground and background separation. In contrast, although FreeSOLO and SAM improve U-Recall, they also raise WI and A-OSE and introduce a slight decrease in K-mAP due to noisy pseudo boxes produced by large models competing with ground-truth annotations. These observations demonstrate that the strength of OW-DAR does not depend on proposal quality but on its ability to accurately identify unknown objects within a large set of mixed foreground and background candidates.

Ablation on reconstruction loss in AREA (Tab. 5) We analyze several representative reconstruction loss functions within the proposed AREA module, including ℓ_1-norm, ℓ_2-norm, and Huber loss (Friedman 2001). Compared to these static methods, our adaptive region-based reconstruction loss achieves the best performance across all metrics, significantly reducing WI and A-OSE. This validates the effectiveness of region-level aggregation and dynamic weighting.

Sensitivity Analysis (Fig. 4) We perform a sensitivity analysis on the masking ratio in the FMR module to assess its impact on detection performance. The optimal results are achieved with a masking ratio of 0.6, striking a strong balance between U-Recall and K-mAP, while minimizing open-set confusion metrics (A-OSE and WI). This configuration introduces moderate perturbation to foreground and background regions, improving semantic reconstruction and foreground-background separability, thereby enhancing the recall of unknown objects.

Visualization

Fig. 3 presents a qualitative comparison of representative methods, including ORTH (Sun, Li, and Mu 2024), MEPU-SS (Fang et al. 2025), SGROD (He et al. 2024), and our OW-DAR. In the first row, ORTH yields sparse predictions and misses multiple unknown objects such as the television and the lamp. MEPU-SS identifies the cat and dog but overlooks several unknown items, including the framed paintings. SGROD incorrectly assigns large background areas (e.g., wall and floor) as unknown, producing many false positives. In contrast, OW-DAR provides accurate and compact detections for both known and unknown objects. In the second row, OW-DAR successfully recovers numerous unknown objects, including the television, tabletop items, and the cabinet. SGROD exhibits redundancy and background confusion, while MEPU-SS and ORTH detect only a subset of small or partially occluded unknown objects, such as the decorative items near the bottom corners. Additionally, OW-DAR reaches the highest inference speed at 35.42 FPS on an RTX 3090, surpassing SGROD (24.51), MEPU-SS (32.25), and ORTH (31.84) while maintaining strong accuracy.

Conclusions

In this paper, we propose OW-DAR, which enhances foreground and background separability through the collaborative design of a fine-grained FMR module and a coarse-grained AREA module. The former guides the model to reconstruct semantically structures, thereby improving its ability to distinguish between foreground and background. The latter focuses on semantically ambiguous boundaries between foreground and background, while simultaneously suppressing the influence of local outliers. Together, these components enable OW-DAR to effectively address the known-class prior bias, achieving strong performance in detecting unknown objects. Importantly, OW-DAR redefines the focus from generating foreground regions to effectively distinguishing unknown objects from a large pool of mixed foreground and background candidates. Therefore, OW-DAR achieves state-of-the-art performance without relying on knowledge from any foundation model. We expect that OW-DAR will promote the broader adoption of OWOD in real-world environments.

Acknowledgments

The authors gratefully acknowledge the support of the National Key Research and Development Program of China (Grant No. 2024YFE0105400).

References

Bai, L.; Huang, Z.; Sun, M.; Cheng, X.; and Cui, L. 2024. Multi-Modal Intelligent Channel Modeling: A New Modeling Paradigm via Synesthesia of Machines. *arXiv preprint arXiv:2411.03711*.

Berthelot, D.; Raffel, C.; Roy, A.; and Goodfellow, I. 2018. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020a. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020b. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Clark, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3): 181–204.

Denouden, T.; Salay, R.; Czarnecki, K.; Abdelzad, V.; Phan, B.; and Vernekar, S. 2018. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*.

Dhamija, A.; Gunther, M.; Ventura, J.; and Boult, T. 2020. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1021–1030.

Doan, T.; Bennani, M. A.; Mazoure, B.; Rabusseau, G.; and Alquier, P. 2021. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, 1072–1080. PMLR.

Doan, T.; Li, X.; Behpour, S.; He, W.; Gou, L.; and Ren, L. 2024. Hyp-ow: Exploiting hierarchical structure learning with hyperbolic distance enhances open world object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1555–1563.

Elakkiya, R.; Teja, K. S. S.; Jegatha Deborah, L.; Bisogni, C.; and Medaglia, C. 2022. Imaging based cervical cancer diagnostics using small object detection-generative adversarial networks. *Multimedia Tools and Applications*, 1–17.

Fang, R.; Pang, G.; Miao, W.; Bai, X.; Zheng, J.; and Ning, X. 2025. Unsupervised recognition of unknown objects for open-world object detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friston, K. 2010. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2): 127–138.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Girshick, R. 2015. Fast r-cnn. *arXiv preprint arXiv:1504.08083*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Gupta, A.; Narayan, S.; Joseph, K.; Khan, S.; Khan, F. S.; and Shah, M. 2022. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9235–9244.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, Y.; Chen, W.; Wang, S.; Liu, T.; and Wang, M. 2024. Recalling Unknowns without Losing Precision: An Effective Solution to Large Model-Guided Open World Object Detection. *IEEE Transactions on Image Processing*.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science*, 313(5786): 504–507.

Jiang, W.; Ge, Y.; Cheng, H.; Chen, M.; Feng, S.; and Wang, C. 2023. Read: Aggregating reconstruction error into out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14910–14918.

Joseph, K.; Khan, S.; Khan, F. S.; and Balasubramanian, V. N. 2021. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5830–5840.

Karri, M.; Annavarapu, C. S. R.; Mallik, S.; Zhao, Z.; and Acharya, U. R. 2022. Multi-class nucleus detection and classification using deep convolutional neural network with enhanced high dimensional dissimilarity translation model on cervical cells. *Biocybernetics and Biomedical Engineering*, 42(3): 797–814.

Khanam, R.; and Hussain, M. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Li, K.; Chen, K.; Wang, H.; Hong, L.; Ye, C.; Han, J.; Chen, Y.; Zhang, W.; Xu, C.; Yeung, D.-Y.; et al. 2022. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *European Conference on Computer Vision*, 406–423. Springer.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Ma, S.; Wang, Y.; Wei, Y.; Fan, J.; Li, T. H.; Liu, H.; and Lv, F. 2023a. Cat: Localization and identification cascade detection transformer for open-world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19681–19690.

Ma, S.; Wang, Y.; Wei, Y.; Fan, J.; Zhang, E.; Sun, X.; and Chen, P. 2023b. SKDF: A Simple Knowledge Distillation Framework for Distilling Open-Vocabulary Knowledge to Open-world Object Detector. *arXiv preprint arXiv:2312.08653*.

Ma, Z.; Yang, Y.; Wang, G.; Xu, X.; Shen, H. T.; and Zhang, M. 2022. Rethinking open-world object detection in autonomous driving scenarios. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1279–1288.

Menezes, A. G.; de Moura, G.; Alves, C.; and de Carvalho, A. C. 2023. Continual object detection: a review of definitions, strategies, and challenges. *Neural networks*, 161: 476–493.

Miller, D.; Nicholson, L.; Dayoub, F.; and Sünderhauf, N. 2018. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 3243–3249. IEEE.

Navon, D. 1977. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3): 353–383.

Nie, W.; Jiao, C.; Chang, R.; Qu, L.; and Liu, A.-A. 2023. CPG3D: Cross-modal priors guided 3D object reconstruction. *IEEE Transactions on Multimedia*, 25: 9383–9396.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Redmon, J. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Sun, Z.; Li, J.; and Mu, Y. 2024. Exploring Orthogonality in Open World Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17302–17312.

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision*, 104: 154–171.

Wang, X.; Yu, Z.; De Mello, S.; Kautz, J.; Anandkumar, A.; Shen, C.; and Alvarez, J. M. 2022. Freesolo: Learning to segment objects without annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14176–14186.

Wang, Y.; Yue, Z.; Hua, X.-S.; and Zhang, H. 2023. Random boxes are open-world object detectors. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6233–6243.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16965–16974.

Zhou, Y. 2022. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7379–7387.

Zohar, O.; Lozano, A.; Goel, S.; Yeung, S.; and Wang, K.-C. 2023. Open world object detection in the era of foundation models. *arXiv preprint arXiv:2312.05745*.

Zohar, O.; Wang, K.-C.; and Yeung, S. 2023. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11444–11453.