OW-OVD: Unified Open World and Open Vocabulary Object Detection

Xing Xi, Yangyang Huang, Ronghua Luo*, Yu Qiu
School of Computer Science & Engineering, South China University of Technology

Abstract

Open world perception expands traditional closed-set frameworks, which assume a predefined set of known categories, to encompass dynamic real-world environments. Open World Object Detection (OWOD) and Open Vocabulary Object Detection (OVD) are two main research directions, each addressing unique challenges in dynamic environments. However, existing studies often focus on only one of these tasks, leaving the combined challenges of OWOD and OVD largely underexplored. In this paper, we propose a novel detector, OW-OVD, which inherits the zero-shot generalization capability of OVD detectors while incorporating the ability to actively detect unknown objects and progressively optimize performance through incremental learning, as seen in OWOD detectors. To achieve this, we start with a standard OVD detector and adapt it for OWOD tasks. For attribute selection, we propose the Visual Similarity Attribute Selection (VSAS) method, which identifies the most generalizable attributes by computing similarity distributions across annotated and unannotated regions. Additionally, to ensure the diversity of attributes, we incorporate a similarity constraint in the iterative process. Finally, to preserve the standard inference process of OVD, we propose the Hybrid Attribute-Uncertainty Fusion (HAUF) method. This method combines attribute similarity with known class uncertainty to infer the likelihood of an object belonging to an unknown class. We validated the effectiveness of OW-OVD through evaluations on two OWOD benchmarks, M-OWODB and S-OWODB. The results demonstrate that OW-OVD outperforms existing state-of-the-art models, achieving a +15.3 improvement in unknown object recall (U-Recall) and a +15.5 increase in unknown class average precision (U-mAP). Our code is available at: https://github.com/xxyzll/OW_OVD.

1. Introduction

Object detection is a fundamental research area in computer vision, with applications in healthcare [8, 27], industrial safety [24, 26], and automation [17, 20].Traditional detectors [39, 60], commonly known as closed-set detectors, require that all categories be present and annotated in the training set. This requirement limits their real-world applicability, where annotating all possible objects is infeasible. To overcome this limitation, two tasks have emerged: Open World Object Detection (OWOD) and Open Vocabulary Object Detection (OVD). As shown in Fig. 1, OVD utilizes a pre-trained text encoder to transform class names into text embeddings, which are subsequently matched with visual embeddings, thereby eliminating the constraint on the number of detectable categories. In contrast, OWOD requires detectors to identify unannotated categories and recognize novel ones through incremental learning. Existing works typically focus on either OWOD or OVD, but each has its limitations. OVD models excel in generalization, enabling zero-shot detection by matching new class names to unseen categories without requiring re-annotation of objects. Nevertheless, they lack the capability to actively discover unknown objects of interest. OWOD models can detect unannotated objects and update through incremental learning but lack OVD's zero-shot capability. To bridge these limitations, we propose OW-OVD, a novel detector designed to effectively address both OVD and OWOD tasks, thus com-

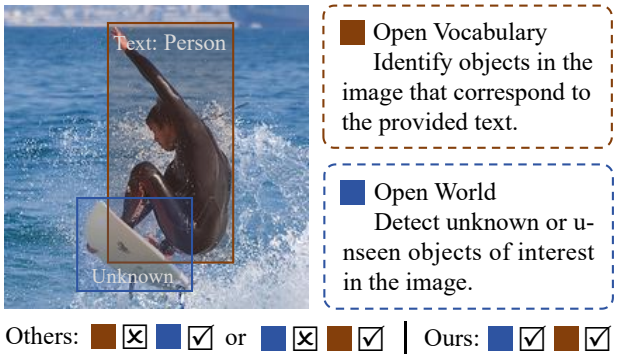

Figure 1. Function description. Previous detectors supported either OVD or OWOD exclusively. Our OW-OVD, however, supports both. It integrates the zero-shot capability of OVD with the proactive unknown object detection and continuous learning features of OWOD, enabling it to recognize new categories.

*Corresponding Author. Ronghua Luo (rhluo@scut.edu.cn)

bining the strengths of both tasks.

To achieve this goal, we build upon the existing OVD detector architecture, augmenting it with the capability to detect unknown objects. Since our aim is to enable the detector to support both OVD and OWOD tasks, existing approaches like FOMO [63], which leverage linear scaling of attribute similarity for predicting unknowns, are deemed unsuitable as they alter the standard inference process of OVD, thereby compromising its inherent zero-shot capability. Therefore, we propose the Visual Similarity Attribute Selection (VSAS) method. First, leveraging standard matching methods in object detection, we categorize the visual embeddings into positive and negative samples. We then compute and aggregate the similarity between all attributes and the visual embeddings. During attribute selection, we derive the probability distribution of attribute similarity. By comparing the similarity distributions of positive and negative samples, we assess the differences between annotated and unannotated regions. Based on these differences, we select attributes that are common between annotated and unannotated regions. Additionally, to prevent the selected attributes from being overly similar, we introduce a similarity constraint. Finally, for unknown object prediction, we propose the Hybrid Attribute-Uncertainty Fusion (HAUF) method. HAUF incorporates known class uncertainty and weighted attribute similarity to estimate the likelihood of a given visual region being classified as unknown.

The unknown prediction of OW-OVD does not alter the inference process of OVD, enabling it to possess both the active discovery of objects of interest in OWOD and the zero-shot detection capability of OVD. Additionally, we validate our approach on two OWOD task benchmarks, M-OWODB and S-OWODB, which are composed of COCO [33] and VOC [9]. The results demonstrate that OW-OVD significantly outperforms previous SOTA models in both known and unknown categories, achieving a notable performance gain of +15.3 in U-Recall and +4.3 in mAP on the S-OWODB task 1. Furthermore, when evaluated using a more stringent metric (U-mAP), OW-OVD achieves a more substantial performance gap (+15.5 U-mAP) compared to SOTA models. We summarize our contributions as follows:

- To the best of our knowledge, we are the first to propose a model that simultaneously possesses the strengths of both OWOD and OVD tasks.
- We propose the Visual Similarity Attribute Selection (VSAS) method that identifies attributes common to both annotated and unannotated regions by comparing the similarity distributions of attributes in these regions.
- We propose the Hybrid Attribute-Uncertainty Fusion (HAUF) method, combining known class uncertainty and weighted attribute similarity to identify unknown objects without altering the OVD inference process.
- We validate OW-OVD on the OWOD benchmark tasks, M-OWODB and S-OWODB. The results show that OW-OVD outperforms existing state-of-the-art (SOTA) models in both known and unknown categories, with improvements of +15.3 in U-Recall and +4.3 in mAP.

2. Related work

2.1. Object detection

Object detection is a core task in computer vision, aiming to identify and locate all objects of interest within an image. Traditional object detection methods typically employ a two-stage detection process [14, 18, 41]. These detectors first generate coarse candidate boxes through region proposals, followed by fine classification and bounding box regression of these candidate boxes. In recent years, single-stage detectors have gradually become mainstream [45, 46, 49]. Unlike two-stage detectors, single-stage detectors perform dense predictions directly on the feature map, thereby improving detection speed. Research hotspots in single-stage methods include optimization of matching strategies [13, 29, 59], alignment of classification and regression tasks [12, 30, 31], and improvements in post-processing algorithms [3, 43, 47]. Additionally, some studies [1, 5, 62] have introduced techniques from natural language processing (NLP) into the object detection field, transforming the detection task into a set prediction problem. However, all these detectors, collectively referred to as closed-set detectors, assume that all objects of interest are annotated in the training set, without considering the emergence of new categories or unannotated objects in real-world scenarios. This limitation affects their applicability in open environments.

2.2. Open vocabulary object detection

Open Vocabulary Object Detection (OVD) treats the detection problem as a region-to-text matching problem [58]. Since the matching process is not constrained by the number of texts, OVD detectors can theoretically detect an unlimited number of categories [52, 61]. Additionally, OVD's text encoders benefit from pre-trained models such as CLIP [40]. Compared to traditional object detection methods, OVD exhibits greater robustness and flexibility in handling open vocabulary and unknown categories [11, 37]. Current research in OVD primarily focuses on large-scale pre-training [23, 28, 56], alignment of text and visual regions [32, 53, 57], and efficient knowledge distillation [15, 36, 48]. These methods allow OVD detectors to effectively detect new categories by merely providing the class names. However, OVD detectors cannot actively discover unknown objects and require all categories to be predefined, which limits their flexibility in real-world applications where new, unannotated objects frequently emerge.

2.3. Open world object detection

Open World Object Detection (OWOD), unlike OVD, treats the identification of unannotated objects as a progressive process [21]. During inference, it actively detects objects that are likely unknown and presents them to annotators. Annotators then select and annotate these unknown objects, adding them as new categories to the dataset. In subsequent incremental learning processes, the detector needs to be fine-tuned on the newly added categories to gradually recognize more objects. Early OWOD methods focused on mining the background during training, marking parts of the background that meet certain criteria as unknown [16, 34, 64]. In recent years, the emergence of large visual models has flourished in this field, such as SAM [25]. Many studies utilize general visual models to detect all objects in an image, filter out known objects, and retain the remaining ones as pseudo-labels, a process known as knowledge distillation [19, 35, 54]. OWOD detectors mimic the continuous optimization process in real-world environments to achieve open-domain perception. However, their training and inference processes are similar to closed-set detectors and lack the zero-shot inference capability of OVD detectors. In this paper, we propose OW-OVD, a detector that supports both OVD and OWOD tasks, combining their advantages.

3. Methodology

3.1. Methodology structure

In Sec. 3.2, we provide an overview of the OVD and OWOD tasks, along with their respective advantages. Fig. 2 illustrates the structure and training process of OW-OVD, including attribute generation (Sec. 3.3) and attribute selection based on visual similarity (Sec. 3.4). Finally, in Sec. 3.5, we propose the hybrid attribute uncertainty fusion method for predicting unknowns. The pseudocode and analysis of our method are presented in Appendix E.

3.2. Task overview and advantage

Open Vocabulary Object Detection formalizes the object detection process as a matching problem between visual regions and textual descriptions. Given an input image I, the visual encoder $f_{vis}(\cdot)$ maps it to a set of visual embeddings $E_{vis} = \{e_{vis_1}, ..., e_{vis_{|E_{vis}|}}\}$. Simultaneously, the text encoder $f_{txt}(\cdot)$ encodes a given set of class labels $C = \{c_1, .., c_j\}$ into corresponding text embeddings $E_C = \{e_{c_1}, ..., e_{c_{|E_C|}}\}$. To determine whether a visual embedding e_{vis_i} belongs to class c_j, the similarity $Sim(e_{vis_i}, e_{c_{|C|}})$ is computed, with the cosine similarity being a commonly used metric; The primary advantage of OVD lies in its zero-shot capability, meaning that when detecting new classes, it is only necessary to input the textual description of the new class into the text encoder, allowing the model to recognize and classify the new category without the need for retraining or additional labeled data.

Open World Object Detection progressively identifies unknown objects through incremental learning. OWOD divides the detection task into a series of subtasks, $T = \{T_1, ..., T_{|T|}\}$. During the training of subtask T_i, the model is introduced to new categories $\mathcal{K}_{T_i}$. At this stage, the detector has already encountered categories $\mathcal{K}_{kn} = \{\mathcal{K}_{T_1}, ..., \mathcal{K}_{T_i}\}$. During inference, the detector is required not only to recognize known categories $\mathcal{K}_{kn}$ but also to predict unannotated objects, labeling them as unknown. In subsequent task T_{i+1}, annotators select new objects from the unknown category and label them to generate $\mathcal{K}_{T_{i+1}}$. The detector is then fine-tuned based on these newly annotated data to detect new categories. The OWOD task simulates the real-world process of gradually annotating new objects and continuously updating the detector; Its significant advantage lies in its ability to proactively identify potential unknown objects and maintain efficient detection capabilities for new categories through incremental learning.

3.3. Attribute generation

To minimize potential biases introduced by individual descriptions of objects, we adopt a similar approach to prior work by employing a large language model (LLM) to generate attributes [63]. Given known category names, we input them into the LLM, prompting it to list relevant features such as color and texture. These features are then inserted into a predefined template, *Object which Feature has/is/etc Response*, to generate descriptive sentences about the objects. These sentences are subsequently encoded into attribute embeddings by the text encoder:

$$E_{att} = f_{txt}(Sen) \in \mathbb{R}^{s \times d}, \tag{1}$$

where f_{txt} represents the text encoder, typically based on the pre-trained CLIP model [40]. Sen refers to the generated sentences, with s being the number of sentences and d representing the dimensionality of the embedding vectors, generally set to 512. Similarly, we encode the class names into class embeddings $E_C \in \mathbb{R}^{k \times d}$, where k denotes the number of known classes.

3.4. Visual similarity attribute selection

To mitigate bias in object descriptions, we utilize a LLM to generate features. However, the features generated by the LLM often exhibit redundancy or overlap. For instance, in Task 1, the LLM generated nearly 2000 features. To reduce the negative impact of these redundant features on model performance, we propose a feature selection method called Visual Similarity Attribute Selection (VSAS). This method selects the most generalizable attributes by calculating the similarity between attributes and both positive and negative

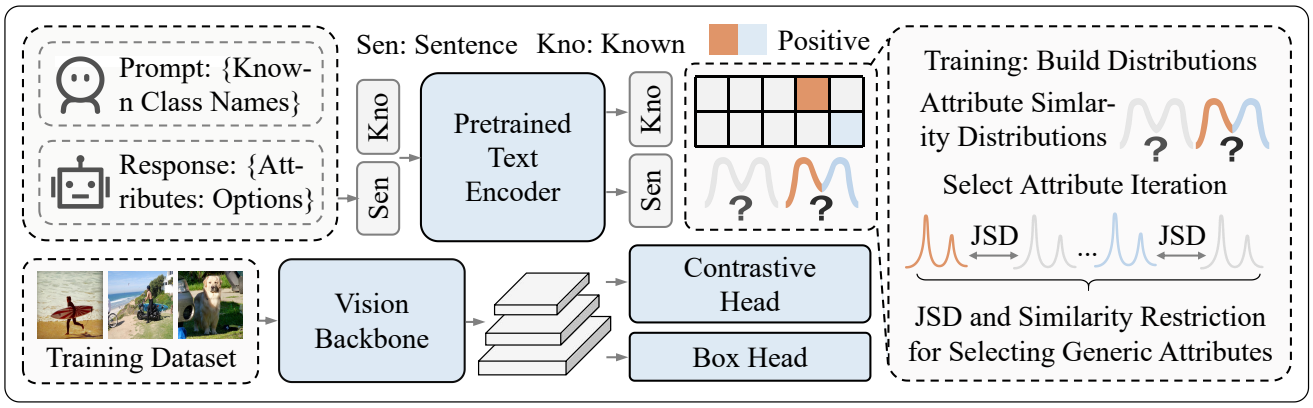

Figure 2. Model structure and training pipeline. Our OW-OVD model is based on the standard OVD detector (YOLO-World [4]), which comprises a text encoder and a visual encoder (vision backbone and feature pyramid). The training process of OW-OVD consists of two primary steps: distribution construction and attribute selection. During training, the detector constructs distributions by measuring the similarities between attributes and both positive and negative samples. In the attribute selection phase, we use Jensen-Shannon Divergence (JSD) and similarity constraints, implemented through an iterative process, to progressively select the most generalizable attributes.

samples during training. VSAS consists of two main steps: distribution construction and iterative attribute selection.

Distribution Construction. To evaluate the differences in each attribute between positive and negative samples, we calculate the similarity of these attributes to the corresponding regions. For each image I in the training set, we first encode it into a visual embedding E_{vis} using the visual encoder. Subsequently, the box and contrastive head decode E_{vis} into bounding boxes and class similarity scores:

$$P_{cls} = h_{cls}(E_{vis}, E_C),\ P_{box} = h_{box}(E_{vis}), \tag{2}$$

where h_{cls} and h_{box} represent the bounding box decoder head (Box Head) and the classification head (Contrastive Head), respectively. E_C denotes the text embeddings obtained from known class names via the text encoder. Subsequently, P_{cls} and P_{box} are matched with all known annotations. Using the matching method's scores, E_{vis} is split into two subsets: positive and negative samples:

$$\begin{aligned} S =& \pi(\lambda\,(L_{cls}(P_{cls}, G_{cls})) + L_{box}(P_{box}, G_{box})), \\ E^+_{vis} =& \{e_{vis_i} | S[i] \geq \alpha\},\ E^-_{vis} = \{e_{vis_i} | S[i] < \alpha\}, \end{aligned} \tag{3}$$

where L_{cls} and L_{box} are the classification and box matching loss functions, respectively. G_{cls} and G_{box} are the annotations of known objects. π is the assignment method, such as TAL[12]. λ represents a weighting factor for combining the two losses. $S[i]$ denotes the matching score for the i-th sample, and α is a threshold used to separate positive (E^+_{vis}) and negative (E^-_{vis}) samples. Subsequently, we compute the similarity of the attributes with respect to E^+_{vis} and E^-_{vis}, in order to record the similarity distribution of the current attribute between the labeled and unlabeled regions:

$$d^+_i = h_{cls}(e^+_{vis}, e_{att_i}),\ d^-_i = h_{cls}(e^-_{vis}, e_{att_i}), \tag{4}$$

where $e^+_{vis} \in E^+_{vis}$, $e^+_{vis} \in E^-_{vis}$ and $e_{att_i} \in E_{att}$. We collected d^+_i and d^-_i across the entire training set to establish frequency distributions. Finally, these distributions are normalized by the total occurrences to generate the probability distributions D^+_i and D^-_i.

Iterative Attribute Selection. Ideally, we aim to obtain accurate descriptions of potential objects. However, as no information (including their names) is available about these objects, we cannot directly use the LLM to generate corresponding descriptions. Therefore, selecting attributes from the descriptions of known categories is a viable solution. We aim to select sentences from these descriptions that can be applied to unknown objects. Since these sentences are generated based on known categories, they must be general enough to describe unknown categories effectively, rather than being specific to known categories. In other words, these sentences should exhibit similarity when describing both annotated (known classes) and unannotated regions (unknown classes). Consequently, we select features most similar to D^+_i and D^-_i and use JSD (Jensen-Shannon Divergence) to evaluate the similarity between the two probability distributions:

$$\begin{aligned} \hat{E}_{att} =& \hat{E}_{att} \cup \underset{e_{att_i}}{Argmin}\ JSD(D^+_i, D^-_i), \\ JSD(D^+_i, D^-_i) =& \frac{1}{2}(KL(D^+_i, M) + KL(D^-_i, M)), \end{aligned} \tag{5}$$

where $\hat{E}_{att}$ represents the selected attributes with the highest generality across both known and unknown categories. $KL(\cdot)$ is the Kullback-Leibler Divergence, $M = \frac{1}{2}(D^+_i + D^-_i)$. Eq. (5) ensures that the most generalizable attributes are identified during the iteration process. To prevent selecting overly similar attributes throughout the iterations, we

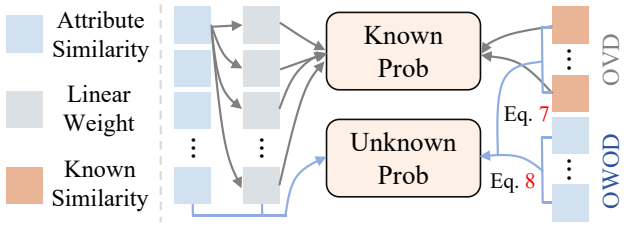

Figure 3. Illustration of the inference process. The left side illustrates the inference process from previous approaches [63], where known and unknown classes are predicted based on a linear combination of attribute similarities. On the right side, our HAUF is presented. In HAUF, the prediction of unknown objects does not interfere with the prediction of known categories, thus preserving the zero-shot detection capability of OVD.

have introduced an additional similarity constraint. Specifically, in each iteration, we penalize features that are too similar to the already selected attributes:

$$\begin{aligned}\hat{E}_{att} = &\hat{E}_{att} \cup \underset{e_{att_i}}{Argmin} \; (\; \beta \cdot JSD(D_i^+, D_i^-) + \\ &(1-\beta) \cdot \frac{1}{|\hat{E}_{att}|} \sum \sigma(sim(e_{att_i}, \hat{E}_{att})) \;),\end{aligned} \tag{6}$$

where $\sigma(\cdot)$ is the Sigmoid function and the sim denotes the cosine similarity. β is a hyperparameter and used to balance their contribution.

3.5. Hybrid attribute-uncertainty fusion

Our objective is to develop a detector that leverages the advantages of both OWOD and OVD. Thus, utilizing attribute-based linear scaling methods from prior work [63] to predict unknown objects is not feasible (Fig. 3 Left). To retain the zero-shot detection capability of the OVD detector, we propose a Hybrid Attribute-Uncertainty Fusion (HAUF) method. HAUF discards the additional linear layers and predicts unknown objects in a parallel manner. Specifically, we consider unknown objects from two perspectives: their similarity to generic attributes and their uncertainty in relation to known categories:

$$\begin{aligned}P_b &= \frac{1}{\gamma} \sum_{\gamma} softmax(s_\gamma) \; s_\gamma, s = h_{cls}(e_{vis}, \hat{E}_{att}), \\ P_{un} &= \frac{1}{k} \sum_{k} -(p_k \; log(p_k) + (1-p_k) log(1-p_k)),\end{aligned} \tag{7}$$

where s_γ is the top γ attribute scores. e_{vis} denotes the visual embedding. P_b is used to distinguish unknown objects from the background, while P_{un} represents the entropy of the model's predictions for the current object relative to known categories. Both perspectives evaluate the likelihood of an object being a positive sample, but relying solely on these could lead to misclassifying known objects as unknown. To prevent this, we combine an out-of-distribution probability for known classes to differentiate them:

$$P_u = \frac{1}{2}(P_b + P_{un})(1 - max(P_C)), \tag{8}$$

where $max(P_C)$ denotes the largest known confidence score. As shown on the right side of Fig. 3, the prediction of unknown objects in HAUF does not interfere with the inference logic for known categories. Consequently, OW-OVD retains the advantages of both OVD and OWOD.

4. Experiments

4.1. Experimental setting

Datasets. Consistent with previous work [34, 51, 64], we combined the MS COCO [33] and VOC [9] datasets and split them into two benchmark datasets, M-OWODB and S-OWODB, following the setups of ORE [21] and OW-DETR [16]. In the M-OWODB benchmark, the mixed dataset is split into four subtasks. Each subtask contains 20 new categories, and the detector is fine-tuned on existing knowledge to adapt to these categories. For more detailed divisions, refer to Appendix A.

Metric. We use the most common evaluation metric in object detection, mean Average Precision (mAP), to assess the performance of the detector on known categories. This metric is also applied to both known and newly introduced categories in subsequent tasks. For unknown categories, we utilize two metrics: U-Recall (recall for unknowns) and U-mAP. Initially, we evaluate all methods using U-Recall. For methods that incorporate large visual models (e.g., SAM), we adopt the stricter metric, U-mAP.

Details. Considering the practical requirements for both accuracy and speed, we selected the YOLO-World detector [4] for our experiments. During training, we constructed distributions in the first epoch. For subsequent tasks, we combine the distributions of previous and current tasks, selecting all attributes. We used the large version of YOLO-World, with a learning rate set to 5e-5. We used two 4090 GPUs (48 GB total), with a batch size of 16 per GPU. In S-OWODB, we set α to 0.75, β to 0.3 and γ to 10. For M-OWODB, we set them to 0.55, 0.2, and 10, respectively. Refer to the Appendix for the impact of hyperparameters. Following previous work [63], we used GPT-3.5 as the LLM and select 25 attributes for each known class, with all code implementations built on mmdetection[2]. For other settings, we follow the same approach as the baseline.

4.2. Comparison of open world object detection

We compared our method with other SOTA methods, using two primary metrics: U-Recall and U-mAP. For additional comparisons, please refer to Appendix C.

U-Recall. Tab. 1 shows a comparison with other SOTA methods on the OWOD standard evaluation bench-

Table 1. Comparison of open-world object detection performance (Recall). The table compares our OW-OVD with previous SOTA methods in recall. The upper part shows M-OWODB, and the lower part shows S-OWODB results. U-Recall measures recall on unknown classes, while mAP reflects known class performance. Current Known, Previously Known, and Both represent categories from past tasks, new categories in the current task, and all seen categories. OW-OVD shows a clear advantage in recall for unknowns and detection for knowns.

Task IDs(→)	**Task 1**		**Task 2**				**Task 3**				**Task 4**		
Method	U-Recall (↑)	mAP (↑) Current Known	U-Recall (↑)	mAP(↑) Previously Known	mAP(↑) Current Known	mAP(↑) Both	U-Recall (↑)	mAP(↑) Previously Known	mAP(↑) Current Known	mAP(↑) Both	mAP(↑) Previously Known	mAP(↑) Current Known	mAP(↑) Both
ORE_EBUI [21]	4.9	56.0	2.9	52.7	26.0	39.4	3.9	38.2	12.7	29.7	29.6	12.4	25.3
OW-DETR [16]	7.5	59.2	6.2	53.6	33.5	42.9	5.7	38.3	15.8	30.8	31.4	17.1	27.8
PROB [64]	19.4	59.5	17.4	55.7	32.2	44.0	19.6	43.0	22.2	36.0	35.7	18.9	31.5
CAT [34]	23.7	60.0	19.1	55.5	32.7	44.1	24.4	42.8	18.7	34.8	34.4	16.6	29.9
RandBox [51]	10.6	61.8	6.3	-	-	45.3	7.8	-	-	39.4	-	-	35.4
ORTH [44]	24.6	61.3	26.3	55.5	38.5	47.0	29.1	46.7	**30.6**	41.3	42.4	24.3	37.9
Hyp-OW [7]	23.5	59.4	20.6	-	-	44.0	26.3	-	-	36.8	-	-	33.6
MEPU-FS [10]	31.6	60.2	30.9	57.3	33.3	44.8	30.1	42.6	21.0	35.4	34.8	19.1	30.9
SGROD [19]	34.3	59.8	32.6	56.0	32.3	44.9	32.7	42.8	22.4	36.0	35.5	18.5	31.2
SKDF [35]	39.0	56.8	36.7	52.3	28.3	40.3	36.1	36.9	16.4	30.1	31.0	14.7	26.9
KTCN [54]	41.5	60.2	38.6	55.8	36.3	46.0	39.7	43.5	22.1	36.4	35.1	16.2	30.4
Ours: OW-OVD	**50.0**	**69.4**	**51.7**	**69.5**	**41.7**	**55.6**	**50.6**	**55.5**	29.8	**47.0**	**47.0**	**25.2**	**41.6**
OW-DETR [16]	5.7	71.5	6.2	62.8	27.5	43.8	6.9	45.2	24.9	38.5	38.2	28.1	33.1
CAT [34]	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 [64]	17.6	73.4	22.3	66.3	36.0	50.4	24.8	47.8	30.4	42.0	42.6	31.7	39.9
ORTH [44]	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
Hyp-OW [7]	23.9	72.7	23.3	-	-	50.6	25.4	-	-	46.2	-	-	44.8
MEPU-FS [10]	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
SGROD [19]	48.0	73.2	48.9	64.7	36.7	50.0	47.7	47.4	32.4	42.4	42.5	32.6	40.0
SKDF [35]	60.9	69.4	60.0	63.8	26.9	44.4	58.6	46.2	28.0	40.1	41.8	29.6	38.7
Ours: OW-OVD	**76.2**	**78.6**	**79.8**	**78.5**	**61.5**	**69.6**	**78.4**	**69.6**	**55.1**	**64.7**	**64.8**	**56.3**	**62.7**

Table 2. Comparison of open-world object detection (U-mAP). The table compares OW-OVD with other SOTA methods using foundation models, focusing on U-mAP for unknown classes.

Method	M-OWODB			S-OWODB		
	Task 1	Task 2	Task 3	Task 1	Task 2	Task 3
MEPU-FS [10]	-	-	-	7.5	5.6	3.4
SGROD [19]	2.4	**4.9**	0.3	1.9	3.9	1.6
SKDF [35]	1.5	1.2	0.7	6.1	4.3	2.0
KTCN [54]	1.2	0.7	0.5	-	-	-
Ours: OW-OVD	**8.6**	4.3	**4.0**	**23.0**	**18.1**	**16.9**

marks M-OWODB and S-OWODB. Consistent with previous methods, we discarded the ORE energy-based model (ORE_EBUI) due to potential data leakage issues. Our OW-OVD showed significant performance advantages. On the M-OWODB benchmark, OW-OVD achieved double the unknown recall compared to ORTH [44] (+25.4). For methods using visual foundation models, we maintained a notable lead. MEPU-FS [10], utilizing FreeSOLO [50] as the self-supervised backbone, was outperformed by OW-OVD with an 18.4 U-Recall advantage. SGROD [19], leveraging GLIP's [28] knowledge, was surpassed by a recall advantage of 15.7. Even for models leveraging SAM's generalization capabilities, OW-OVD remained ahead (+SKDF [35] 10, +KTCN [54] 8.5). On the S-OWODB benchmark, the performance gap further widened; for instance, our method exceeded SKDF by 16.3. Additionally, in known class performance, our method exhibited advantages, surpassing KTCN by 11.2 AP in Task 4.

U-mAP. When applying visual foundation models, the gap between unknown and known recalls narrows. Hence, we adopted a more stringent metric, AP, to compare methods using visual base models. The results are shown in Tab. 2. Other methods exhibited poor detection accuracy, resulting in lower AP scores. In contrast, our OW-OVD showed a significant AP advantage. In the M-OWODB benchmark, we achieved a +6.2 AP advantage (8.6 vs 2.4). In the S-OWODB benchmark, OW-OVD achieved an AP of 23.0, while MEPU-FS achieved only 7.5, demonstrating a significant performance gap (+15.5 AP).

Table 3. Comparison of incremental object detection performance. The table compares the performance of OW-OVD with other SOTA methods in incremental learning on the PASCAL VOC dataset [9]. The table shows three category splits: 10+10, 15+5, and 19+1. The grey areas indicate the new categories introduced in the second task. mAP represents the mean average precision at the end of all training tasks. Our method demonstrates significant performance advantages across all category splits.

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 [42]	69.9	70.4	69.4	54.3	48.0	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 [38]	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 [21]	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	77.0	67.7	64.5
Meta-ILOD [22]	76.0	74.6	67.5	55.9	57.6	75.1	85.4	77.0	43.7	70.8	60.1	66.4	76.0	72.6	74.6	39.7	64.0	60.2	68.5	60.5	66.3
ROSETTA [55]	74.2	76.2	64.9	54.4	57.4	76.1	84.4	68.8	52.4	67.0	62.9	63.3	79.8	72.8	78.1	40.1	62.3	61.2	72.4	66.8	66.8
OW-DETR [16]	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 [64]	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
CAT [34]	76.5	75.7	67.0	51.0	62.4	73.2	82.3	83.7	42.7	64.4	56.8	74.1	75.8	79.2	78.1	39.9	65.1	59.6	78.4	67.4	67.7
ORTH [44]	82.4	77.3	78.2	59.7	61.2	84.3	90.1	80.2	49.8	81.7	58.2	74.0	82.9	81.0	81.2	38.3	70.8	68.0	77.4	70.2	72.3
Ours: OW-OVD	97.3	96.1	90.9	73.4	77.1	95.7	92.3	95.3	75.2	89.3	79.2	94.0	97.3	94.9	91.5	53.7	90.3	82.5	95.3	82.3	**87.2**
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 [42]	70.5	79.2	68.8	59.1	53.2	75.4	79.4	78.8	46.6	59.4	59.0	75.8	71.8	78.6	69.6	33.7	61.5	63.1	71.7	62.2	65.8
Faster ILOD [38]	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 [21]	75.4	81.0	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
Meta-ILOD [22]	78.4	79.7	66.9	54.8	56.2	77.7	84.6	79.1	47.7	75.0	61.8	74.7	81.6	77.5	80.2	37.8	58.0	54.6	73.0	56.1	67.8
ROSETTA [55]	76.5	77.5	65.1	56.0	60.0	78.3	85.5	78.7	49.5	68.2	67.4	71.2	83.9	75.7	82.0	43.0	60.6	64.1	72.8	67.4	69.2
OW-DETR [16]	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 [64]	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
CAT [34]	75.3	81.0	84.4	64.5	56.6	74.4	84.1	86.6	53.0	70.1	72.4	83.4	85.5	81.6	81.0	32.0	58.6	60.7	81.6	63.5	72.2
ORTH [44]	82.7	80.4	78.5	55.3	65.5	81.0	89.8	85.9	52.6	84.6	62.3	78.4	82.7	81.1	84.2	46.5	71.6	79.0	82.5	79.2	74.7
Ours: OW-OVD	97.1	96.0	90.6	74.5	76.8	95.7	92.4	95.0	74.7	87.6	79.6	94.4	97.4	94.4	91.5	54.9	89.0	81.4	95.3	79.3	**86.9**
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 [42]	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 [38]	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 [21]	67.3	76.8	60.0	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
Meta-ILOD [22]	78.2	77.5	69.4	55.0	56.0	78.4	84.2	79.2	46.6	79.0	63.2	78.5	82.7	79.1	79.9	44.1	73.2	66.3	76.4	57.6	70.2
ROSETTA [55]	75.3	77.9	65.3	56.2	55.3	79.6	84.6	72.9	49.2	73.7	68.3	71.0	78.9	77.7	80.7	44.0	69.6	68.5	76.1	68.3	69.6
OW-DETR [16]	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 [64]	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
CAT [34]	86.0	85.8	78.8	65.3	61.3	71.4	84.8	84.8	52.9	78.4	71.6	82.7	83.8	81.2	80.7	43.7	75.9	58.5	85.2	61.1	73.8
ORTH [44]	83.8	84.7	77.0	62.9	60.8	80.9	88.6	85.8	51.1	81.4	67.2	86.7	86.3	83.4	83.4	44.7	74.5	73.1	81.1	74.9	75.6
Ours: OW-OVD	97.0	96.0	90.8	75.1	76.3	95.3	92.4	95.0	75.0	86.3	78.9	94.3	97.2	94.3	91.6	52.1	90.0	82.2	95.4	80.4	**86.8**

4.3. Comparison of incremental object detection

Tab. 3 presents a detailed comparison of incremental learning methods. Other methods underperformed in incremental learning tasks. For instance, in the 10+10 split, ORTH [44] achieved only 72.3 AP. Our OW-OVD retains the zero-shot capability of OVD and follows the same prompt fine-tuning process during training. In the 10+10 setup, OW-OVD achieved 87.2 AP, surpassing ORTH by 14.9. Similarly, in the 15+5 and 19+1 setups, we led with 86.9 AP and 86.8 AP, outperforming the previous best models by 12.2 and 11.2, respectively. Additionally, OW-OVD shows consistent performance across different experimental splits. The performance gap between the 10+10 and 19+1 setups is only 0.4 AP. In contrast, other methods are highly sensitive to the number of known classes. For example, PROB [64] achieved 66.5 AP in the 10+10 setup and 72.6 AP in the 19+1 setup, with a gap of 6.1. This sensitivity hinders the detector's applicability in open-world environments, which often contain a large number of unlabelled objects. Our OW-OVD overcomes this limitation by remaining insensitive to the number of unknown classes.

4.4. Visualization

Fig. 4 shows the visual comparison with other methods. OW-OVD predicts unknown classes more accurately. KTCN and SGROD, using SAM, predicted many unknown objects, but often incompletely or as object parts, e.g., KTCN predicted both a lamp and its base as separate unknowns. They also produced duplicate predictions. MEPU and SKDF had similar issues, misidentifying parts of objects, like MEPU with a bicycle and SKDF with a toilet. OW-OVD, however, predicted whole objects without inter-

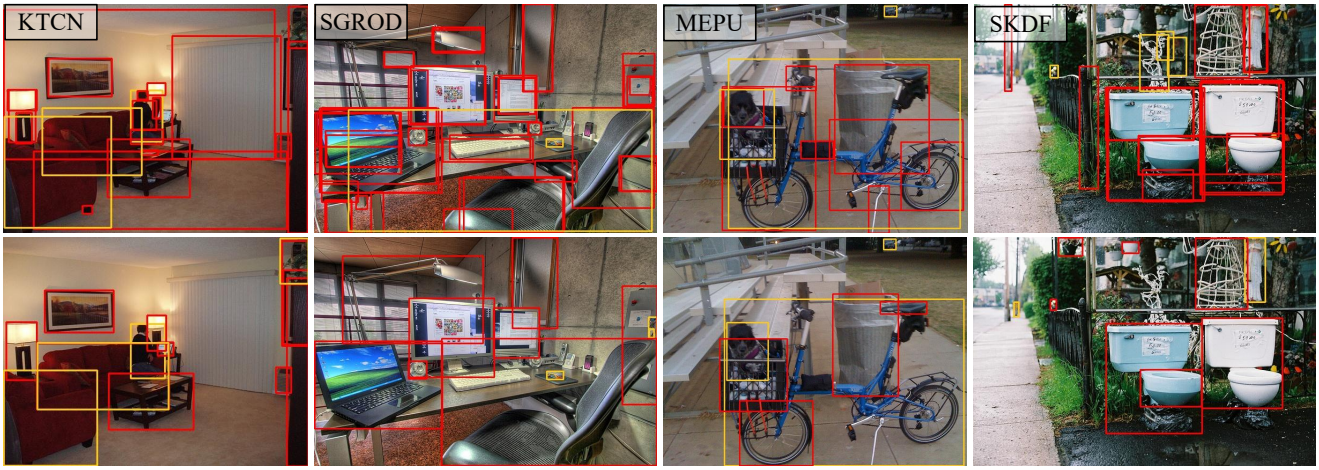

Figure 4. Visualization. The first row displays the results of other state-of-the-art (SOTA) methods that use foundation models, while the second row presents the outcomes of our OW-OVD approach. Red indicates unknown classes, and yellow represents known objects. OW-OVD demonstrates superior performance in unknown object detection, particularly in terms of detection accuracy.

Table 4. Ablation studies. We show the impact of all configurations. Base+GT: all unknown names. All attr: all attributes. OOD prob: out-of-distribution probability (Eq. (8)). Attr sel: attribute selection (Eq. (5)). Sim restr: similarity restriction (Eq. (6)). Known uncer: known class uncertainty (P_{un}). Top attr: most similar attribute (P_b). For hyperparameter, refer to Appendix B.

Method	M-OWODB			S-OWODB		
	U-mAP	U-Recall	mAP	U-mAP	U-Recall	mAP
Base+GT	23.8	58.4	69.0	54.6	86.2	76.9
All attr	4.3	54.4	69.1	15.0	80.9	78.0
OOD prob	5.9	54.4	69.1	18.1	80.9	78
Attr sel	6.9	54.3	69.2	20.5	81.2	78.4
Sim restr	7.1	54.8	69.2	20.7	81.5	78.4
Known uncer	7.4	53.5	69.3	20.7	80.3	78.5
Top attr	8.6	50.0	69.4	23.0	76.5	78.6

nal redundancies, resulting in more accurate outcomes.

4.5. Ablation studies

Tab. 4 shows the performance improvement of each configuration. For additional experimental details (e.g., hyperparameter settings), refer to the Appendix. Firstly, we used all unknown class names to test the baseline model's performance upper limit (Base+GT), achieving 23.8 U-mAP in M-OWODB and 54.6 U-mAP in S-OWODB. Next, using all attributes for unknown object prediction resulted in only 4.3 U-mAP for M-OWODB. Introducing OOD probability (Eq. (8)) increased the U-mAP to 5.9 for M-OWODB and to 18.1 for S-OWODB, indicating that OOD enhances the detector's ability to distinguish known from unknown. With attribute selection (Attr sel, Eq. (5)), performance increased to 6.9 U-mAP for M-OWODB and 20.5 U-mAP for S-OWODB, demonstrating the necessity of attribute selection. Despite using fewer attributes, the detector achieved higher performance. Introducing similarity restriction (Sim restr, Eq. (6)) led to improvements in both U-mAP and U-Recall, suggesting that similarity restriction helps select more diverse general features, a result not achieved by other methods. Finally, using top attribute weighting (Top attr), U-mAP improved to 8.6 for M-OWODB and 23.0 for S-OWODB, significantly surpassing the initial 4.3 and 15 U-mAP. Additionally, OW-OVD maintained or even enhanced performance on known classes, with mAP increasing from 69.0 to 69.4 and from 78.0 to 78.6.

5. Conclusion

In this paper, we introduce OW-OVD, a detector that combines open vocabulary and open world object detection tasks. To achieve this, we propose two methods: visual similarity attribute selection and hybrid attribute-uncertainty fusion. The former selects the most general visual attributes, while the latter predicts unknown objects and preserves OVD's zero-shot capability. OW-OVD shows strong performance, particularly in detecting unknown objects. Additionally, OVD's zero-shot capability gives OW-OVD a significant advantage in incremental learning, making it less sensitive to changes in the number of unknown classes. We anticipate that OW-OVD will promote the widespread adoption of detectors in real-world environments.

Acknowledgments

This work was supported by National Key Research and Development Program of China (Grant No. 2024YFE0105400.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[3] Yiqun Chen, Qiang Chen, Qinghao Hu, and Jian Cheng. Date: Dual assignment for end-to-end fully convolutional object detection. *arXiv preprint arXiv:2211.13859*, 2022.

[4] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024.

[5] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Thang Doan, Xin Li, Sima Behpour, Wenbin He, Liang Gou, and Liu Ren. Hyp-ow: Exploiting hierarchical structure learning with hyperbolic distance enhances open world object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1555–1563, 2024.

[8] R Elakkiya, V Subramaniyaswamy, V Vijayakumar, and Aniket Mahanti. Cervical cancer diagnostics healthcare system using hybrid object detection adversarial networks. *IEEE Journal of Biomedical and Health Informatics*, 26(4): 1464–1471, 2021.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[10] Ruohuan Fang, Guansong Pang, Lei Zhou, Xiao Bai, and Jin Zheng. Unsupervised recognition of unknown objects for open-world object detection. *arXiv preprint arXiv:2308.16527*, 2023.

[11] Ruohuan Fang, Guansong Pang, and Xiao Bai. Simple image-level classification improves open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1716–1725, 2024.

[12] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499. IEEE Computer Society, 2021.

[13] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 303–312, 2021.

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

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[16] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9235–9244, 2022.

[17] Vishal Gupta and Monish Gupta. Automated object detection system in marine environment. In *Mobile Radio Communications and 5G Networks: Proceedings of MRCN 2020*, pages 225–235. Springer, 2021.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[19] Yulin He, Wei Chen, Siqi Wang, Tianrui Liu, and Meng Wang. Recalling unknowns without losing precision: An effective solution to large model-guided open world object detection. *IEEE Transactions on Image Processing*, 2024.

[20] Muhammad Ilyas, Hui Ying Khaw, Nithish Muthuchamy Selvaraj, Yuxin Jin, Xinge Zhao, and Chien Chern Cheah. Robot-assisted object detection for construction automation: Data and information-driven approach. *IEEE/Asme Transactions on Mechatronics*, 26(6):2845–2856, 2021.

[21] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5840, 2021.

[22] KJ Joseph, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9209–9216, 2021.

[23] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Detection-oriented image-text pretraining for open-vocabulary detection. *arXiv preprint arXiv:2310.00161*, 2023.

[24] Siyeon Kim, Seok Hwan Hong, Hyodong Kim, Meesung Lee, and Sungjoo Hwang. Small object detection (sod) system for comprehensive construction site safety monitoring. *Automation in Construction*, 156:105103, 2023.

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[26] Min-Chul Kong, Myung-Il Roh, Ki-Su Kim, Jeongyoul Lee, Jongoh Kim, and Gapheon Lee. Object detection method for ship safety plans using deep learning. *Ocean Engineering*, 246:110587, 2022.

[27] Louis Lecrosnier, Redouane Khemmar, Nicolas Ragot, Benoit Decoux, Romain Rossi, Naceur Kefi, and Jean-Yves

Ertaud. Deep learning-based object detection, localisation and tracking for smart wheelchair healthcare mobility. *International journal of environmental research and public health*, 18(1):91, 2021.

[28] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[29] Shuai Li, Chenhang He, Ruihuang Li, and Lei Zhang. A dual weighting label assignment scheme for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9387–9396, 2022.

[30] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020.

[31] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2021.

[32] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[34] Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Thomas H Li, Hongli Liu, and Fanbing Lv. 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*, pages 19681–19690, 2023.

[35] Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Xinyu Sun, Peihao Chen, and Enming Zhang. A simple knowledge distillation framework for open-world object detection. *arXiv preprint arXiv:2312.08653*, 2023.

[36] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022.

[37] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *Pattern recognition letters*, 140:109–115, 2020.

[39] Yifan Pu, Weicong Liang, Yiduo Hao, Yuhui Yuan, Yukang Yang, Chao Zhang, Han Hu, and Gao Huang. Rank-detr for high quality object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[42] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*, pages 3400–3409, 2017.

[43] Peize Sun, Yi Jiang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Onenet: Towards end-to-end one-stage object detection. *arXiv preprint arXiv:2012.05780*, 3, 2020.

[44] Zhicheng Sun, Jinghan Li, and Yadong Mu. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17302–17312, 2024.

[45] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[46] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE transactions on pattern analysis and machine intelligence*, 44(4): 1922–1933, 2020.

[47] Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. End-to-end object detection with fully convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15849–15858, 2021.

[48] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2023.

[49] Ning Wang, Yang Gao, Hao Chen, Peng Wang, Zhi Tian, Chunhua Shen, and Yanning Zhang. Nas-fcos: Fast neural architecture search for object detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11943–11951, 2020.

[50] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14176–14186, 2022.

[51] Yanghao Wang, Zhongqi Yue, Xian-Sheng Hua, and Hanwang Zhang. Random boxes are open-world object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6233–6243, 2023.

[52] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting ev-

erything in the open world: Towards universal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11443, 2023.

[53] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15254–15264, 2023.

[54] Xing Xi, Yangyang Huang, Jinhao Lin, and Ronghua Luo. Ktcn: Enhancing open-world object detection with knowledge transfer and class-awareness neutralization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 1462–1470. International Joint Conferences on Artificial Intelligence Organization, 2024. Main Track.

[55] Binbin Yang, Xinchi Deng, Han Shi, Changlin Li, Gengwei Zhang, Hang Xu, Shen Zhao, Liang Lin, and Xiaodan Liang. Continual object detection via prototypical task correlation guided gating mechanism. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9255–9264, 2022.

[56] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022.

[57] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023.

[58] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[59] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020.

[60] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[61] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.

[62] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[63] Orr Zohar, Alejandro Lozano, Shelly Goel, Serena Yeung, and Kuan-Chieh Wang. Open world object detection in the era of foundation models. *arXiv preprint arXiv:2312.05745*, 2023.

[64] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2023.