PROB: Probabilistic Objectness for Open World Object Detection

Orr Zohar, Kuan-Chieh Wang, Serena Yeung
Stanford University
{orrzohar, wangkua1, syyeung}@stanford.edu

Abstract

Open World Object Detection (OWOD) is a new and challenging computer vision task that bridges the gap between classic object detection (OD) benchmarks and object detection in the real world. In addition to detecting and classifying seen/labeled *objects, OWOD algorithms are expected to detect* novel/unknown *objects - which can be classified and incrementally learned. In standard OD, object proposals not overlapping with a labeled object are automatically classified as background. Therefore, simply applying OD methods to OWOD fails as unknown objects would be predicted as background. The challenge of detecting unknown objects stems from the lack of supervision in distinguishing unknown objects and background object proposals. Previous OWOD methods have attempted to overcome this issue by generating supervision using pseudo-labeling - however, unknown object detection has remained low. Probabilistic/generative models may provide a solution for this challenge. Herein, we introduce a novel probabilistic framework for objectness estimation, where we alternate between probability distribution estimation and objectness likelihood maximization of known objects in the embedded feature space - ultimately allowing us to estimate the objectness probability of different proposals. The resulting **Pr**obabilistic **Ob**jectness transformer-based open-world detector, PROB, integrates our framework into traditional object detection models, adapting them for the open-world setting. Comprehensive experiments on OWOD benchmarks show that PROB outperforms all existing OWOD methods in both unknown object detection ($\sim$ 2$\times$ unknown recall) and known object detection ($\sim$ 10% mAP). Our code is available at https://github.com/orrzohar/PROB.*

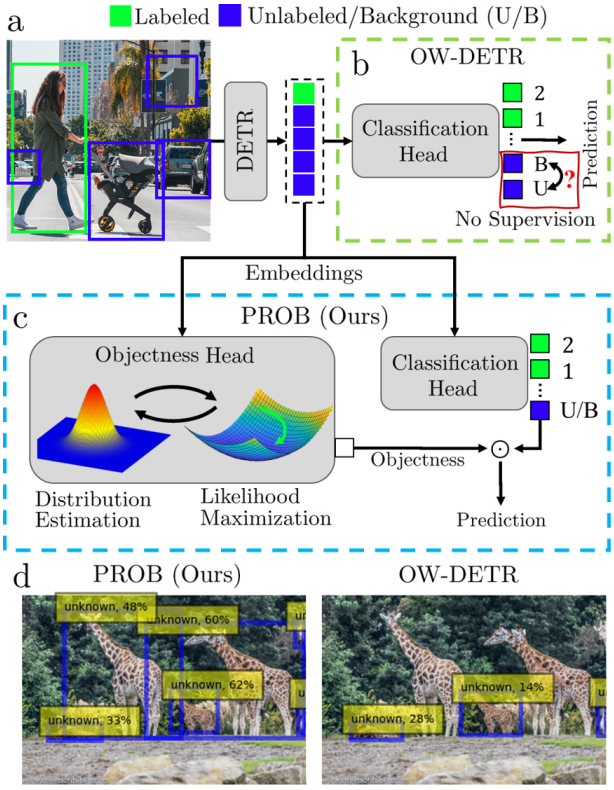

Figure 1. Comparison of PROB with other open world object detectors. (a) Query embeddings are extracted from an image via the deformable DETR model. (b) other open-world detectors attempt to directly distinguish between unlabeled 'hidden' objects and background without supervision (red). (c) PROB's scheme of probabilistic objectness training and revised inference, which performs alternating optimization of (i) Embeddings distribution estimation and (ii) likelihood maximization of embeddings that represent known objects. (d) Qualitative examples of the improved unknown object detection of PROB on the MS-COCO test set.

1. Introduction

Object detection (OD) is a fundamental computer vision task that has a myriad of real-world applications, from autonomous driving [18, 25], robotics [4, 32] to healthcare [6, 12]. However, like many other machine learning systems, generalization beyond the training distribution remains challenging [5] and limits the applicability of existing OD systems. To facilitate the development of machine learning methods that maintain their robustness in the real world, a new paradigm of learning was developed – *Open World Learning* (OWL) [8–10, 16, 17, 21, 27, 29–31, 34]. In OWL, a machine learning system is tasked with reason-

ing about both known and unknown concepts, while slowly learning over time from a non-stationary data stream. In Open World Object Detection (OWOD), a model is expected to detect all previously learned objects while simultaneously being capable of detecting novel *unknown* objects. These flagged unknown objects can be sent to an oracle (human annotator), which labels the objects of interest. The model is then expected to update itself without catastrophically forgetting previous object classes [10].

While unknown object detection is pivotal to the OWOD objective, existing OWOD methods have very low unknown object recall ($\sim$10%) [8, 10, 30, 34]. As such, it is clear that the field has much to improve to meet its actual goal. The difficulty of unknown object detection stems from a lack of supervision as, unlike known objects, unknown objects are not labeled. Hence, while training OD models, object proposals that include an unknown object would be incorrectly penalized as background. Thus far, most OWOD methods have attempted to overcome this challenge by using different heuristics to differentiate between unknown objects and background during training. For example, OW-DETR [8] uses a pseudo-labeling scheme where image patches with high backbone feature activation are determined to be unknown objects, and these pseudo-labels are used to supervise the OD model. In contrast, instead of reasoning about known and unknown objects separately using labels and pseudo-labels, we take a more direct approach. We aim to learn a probabilistic model for general "objectness" (see Fig. 1). Any object – both known and unknown – should have general features that distinguish them from the background, and the learned objectness can help improve both unknown and known object detection.

Herein, we introduce the Probabilistic Objectness Open World Detection Transformer, PROB. PROB incorporates a novel probabilistic objectness head into the standard deformable DETR (D-DETR) model. During training, we alternate between estimating the objectness probability distribution and maximizing the likelihood of known objects. Unlike a classification head, this approach does not require negative examples and therefore does not suffer from the confusion of background and unknown objects. During inference, we use the estimated objectness distribution to estimate the likelihood that each object proposal is indeed an object (see Fig. 1). The resulting model is *simple* and achieves *state-of-the-art* open-world performance. **We summarize our contributions as follows:**

- We introduce PROB - a novel OWOD method. PROB incorporates a probabilistic objectness prediction head that is jointly optimized as a density model of the image features along with the rest of the transformer network. We utilize the objectness head to improve both critical components of OWOD: unknown object detection and incremental learning.
- We show extensive experiments on all OWOD benchmarks demonstrating the PROB's capabilities, which outperform all existing OWOD models. On MS-COCO, PROB achieves relative gains of 100-300% in terms of unknown recall over all existing OWOD methods while improving known object detection performance $\sim 10\%$ across all tasks.
- We show separate experiments for incremental learning tasks where PROB outperformed both OWOD baselines and baseline incremental learning methods.

2. Related Works

Open World Object Detection. The Open World Object Detection task, recently introduced by Joseph *et al*. [10], has already garnered much attention [8, 18, 25, 29–31, 34] due to its possible real-world impact. In their work, Joseph *et al*. [10] introduced ORE, which adapted the faster-RCNN model with feature-space contrastive clustering, an RPN-based unknown detector, and an Energy Based Unknown Identifier (EBUI) for the OWOD objective. Yu *et al*. [31] attempted to extend ORE by minimizing the overlapping distributions of the known and unknown classes in the embeddings feature-space by setting the number of feature clusters to the number of classes, and showed reduced confusion between known and unknown objects. Meanwhile, Wu *et al*. [29] attempted to extend ORE by introducing a second, localization-based objectness detection head (introduced by Kim *et al*. [11]), and reported gains in unknown object recall, motivating objectness's utility in OWOD.

Transformer-based methods have recently shown great potential in the OWOD objective when Gupta *et al*. [8] adapted the deformable DETR model for the open world objective - and introduced OW-DETR. OW-DETR uses a pseudo-labeling scheme to supervise unknown object detection, where unmatched object proposals with high backbone activation are selected as unknown objects. Maaz *et al*. [19] reported on the high class-agnostic object detection capabilities of Multi-modal Vision Transformers (MViTs). They proceeded to utilize MViTs in the supervision of ORE's unknown object detection and reported significant ($\sim 4\times$) gains in its performance. While Maaz *et al*.'s work focused on class-agnostic object detection and did not introduce an OWOD method, their work motivates the possible generalization potential of MViTs and transformer-based models. Recent work in OWOD motivates the use of transformer-based models [8] and the integration of objectness [29] for robust OWOD performance. While previous methods attempted to use objectness estimation [8, 29], none directly integrated it into the class prediction itself. Unlike previous works, we both introduce a novel method for probabilistically estimating objectness and directly integrate it into the class prediction itself, improving unknown object detection.

Class Agnostic Object Detection. Class agnostic object detection (CA-OD) attempts to learn general objectness features given a limited number of labeled object classes. These general features are then used to detect previously unseen object classes. CA-OD methods are expected to localize objects in a class-agnostic fashion. Current SOTA objectness detection [11, 23] all address the same issue; datasets are not densely labeled, and therefore one cannot simply decide that a proposed detection is wrong if it does not overlap with any ground truth label. Saito *et al*. [23] addressed this issue by introducing a custom image augmentation method, BackErase, which pastes annotated objects on an object-free background. Kim *et al*. [11] explored the effect of different losses on learning open-world proposals and found that replacing classification with localization losses, which do not penalize false positives, improves performance. Unfortunately, the direct integration of CA-OD methods has shown poor OWOD performance. For example, the direct integration of Kim *et al*.'s [11] localization-based objectness method into ORE, as presented by Wu *et al*. [29], resulted in a 70% drop in unknown object recall. Although indirectly, our work integrates insights from CA-OD, e.g., the lack of penalization of false positives.

3. Background

Problem Formulation. Let us begin by introducing the notations for standard object detection before extending them to the open-world objective. During training, a model f is trained on a dataset $\mathcal{D} = \{\mathcal{I}, \mathcal{Y}\}$, which contains K known object classes. The dataset contains N images and corresponding labels, $\mathcal{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_N\}$ and $\mathcal{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$, respectively. Each label $\boldsymbol{Y}_i, i \in [1, 2, ..., N]$ is composed of J annotated objects $\boldsymbol{Y}_i = \{\boldsymbol{y}_1, ...\boldsymbol{y}_J\} \in \mathcal{Y}$, which is a set of *object* labels, each of which is a vector containing bounding box coordinates and object class label, i.e., $\boldsymbol{y}_j = [\boldsymbol{l}_j, x_j, y_j, w_j, h_j]$ where $l_j \in \{0, 1\}^K$ is a one-hot vector.

Let us now extend this formulation to the open-world objective. We follow the formulation introduced by Joseph *et al*. [10]. Given a task/time t, there are K^t known in-distribution classes, and an associated dataset $\mathcal{D}^t = \{\mathcal{I}^t, \mathcal{Y}^t\}$, which contains N^t images and corresponding labels. Unlike before, the object class label is now a $K^t + 1$ - dimensional vector $l_j \in \{0, 1\}^{K^t+1}$, where the first element is used to represent unknown objects. There may be an unbounded number of unknown classes, but of these, U^t are classes of interest (unknown classes we would like detect). The model then sends the discovered unknown object objects to an oracle (e.g., a human annotator), which will label the new objects of interest. These newly labeled objects are then used to produce $\mathcal{D}^{t+1}$ (which only contains instances of the U^t newly introduced object classes). The model is then updated, given only $\mathcal{D}^{t+1}$, f^t, and a limited subset of $\mathcal{D}^i, i \in \{0, 1, \ldots, t\}$ to produce f^{t+1} that can detect $K^{t+1} = K^t + U^t$ object classes. This cycle may be repeated as much as needed.

DETR for Open World Learning. DETR-type models [1, 35] have transformed the object detection field due to their simplified design, which has less inductive bias. These models utilize a transformer encoder-decoder to directly transform spatially-distributed features, encoded using some backbone network, into a set of N_{query} object predictions (which can include background predictions). The decoder utilizes N_{query} learned query vectors, each of which queries the encoded image and outputs a corresponding *query embedding*, $\boldsymbol{q} \in \mathbb{R}^D$, i.e., $\boldsymbol{Q} = f^t_{\text{feat}}(I) \in \mathbb{R}^{N_{\text{query}} \times D}$. Each query embedding is then input into the bounding box regression (f^t_{bbox}) and classification (f^t_{cls}) heads (see Fig. 2, bottom). The classification head takes each query embedding and predicts whether it belongs to one of the known objects *or* background/unknown object. The extension of this formulation to open-world object detection is non-trivial, as the model needs to further separate the background/unknown objects from each other, which is unsupervised (the unknown objects are not labeled). To solve this, OW-DETR [8] incorporated an attention-driven pseudo-labeling scheme where the unmatched queries were scored by the average backbone activation, and the top u_k(=5) of them were selected as unknown objects. These pseudo-labels were used during training to supervise unknown object detection, with the inference remaining unchanged.

4. Method

We propose PROB, which adapts the deformable DETR (D-DETR) [35] model for the open world by incorporating our novel 'probabilistic objectness' head. In Sec. 4.1, we describe how the objectness head is trained and used in inference. In Sec. 4.2, we describe how the learned objectness is incorporated in incremental learning; namely, how to learn about new classes in a new task without forgetting old classes. Fig. 2 illustrates the proposed probabilistic objectness open-world object detection transformer, PROB.

4.1. Probabilistic Objectness

The standard D-DETR produces a set of N_{query} query embedding for every image, each of which is used by the detection heads to produce the final predictions. The extension of D-DETR to the open world objective requires the addition of another class label, "Unknown Object". However, unlike the other objects, unknown objects are not labeled – and therefore, one cannot distinguish between them and background predictions while training. Therefore, most OWOD methods attempt to identify these unknown objects and assign them pseudo-labels during training. Rather than

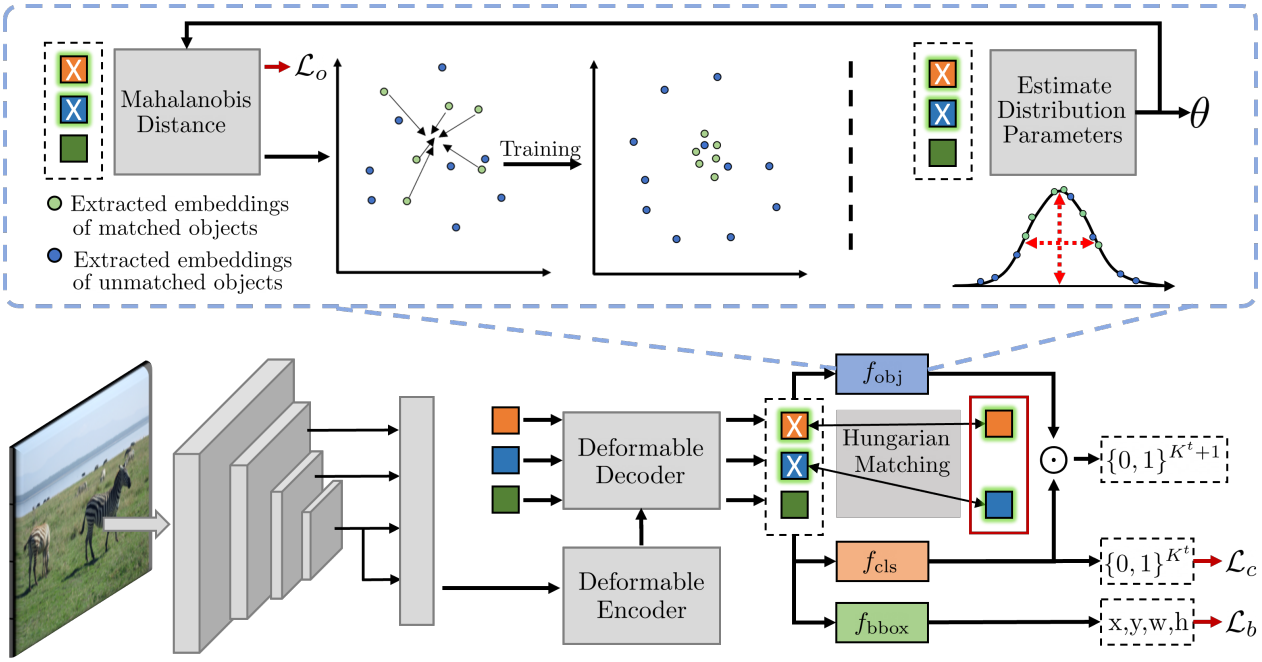

Figure 2. **Overview of the proposed PROB for open-world object detection**. **(top) Probabilistic objectness module.** The probability distribution parameters of all of the query embeddings, θ, are first estimated via the exponential moving average of the mean and covariance estimators. The Mahalanobis distance is then calculated, and the sum of *matched* query embeddings (green dots in scatterplots and white 'X' on query embeddings) are penalized ($\mathcal{L}_o$). This causes the query embeddings of objects to slowly migrate towards the mean, i.e., increased likelihood. **(bottom) Overview of the entire method.** The base architecture of PROB is the deformable DETR (D-DETR) model. Query embeddings are produced by the D-DETR model and subsequently used by the classification, bounding box, and objectness heads. The classification head is trained using a sigmoid focal loss ($\mathcal{L}_c$), while the bounding box head is trained with L1 and gIoU losses ($\mathcal{L}_b$). For class prediction, the learned objectness probability multiplies the classification probabilities to produce the final class predictions.

directly attempting to identify unknown objects, we propose to separate the object (o) and object class ($\boldsymbol{l}|o$) predictions. By separately learning about objectness, $p(o|\boldsymbol{q})$ and object class probability $p(\boldsymbol{l}|o,\boldsymbol{q})$, we no longer need to identify unknown objects while training. The modified inference:

$$\begin{aligned} p(\boldsymbol{l}|\boldsymbol{q}) &= \sum_{\text{i=1,0}} p(\boldsymbol{l}|o=i,\boldsymbol{q}) \cdot p(o=i|\boldsymbol{q}) \\ &= p(\boldsymbol{l}|o=1,\boldsymbol{q}) \cdot p(o=1|\boldsymbol{q}). \end{aligned} \tag{1}$$

As $p(\boldsymbol{l}|o=0,\boldsymbol{q})=0$. The classification head, $f^t_{\text{cls}}(\boldsymbol{q})$ can now operate under the assumption that it already knows if a query embedding represents an object or not, and it learns to imitate $p(\boldsymbol{l}|o,\boldsymbol{q})$. Meanwhile, our objectness head (introduced below) learns to estimate $p(o|\boldsymbol{q})$. Our final class prediction, in the notations of our modified D-DETR model:

$$p(\boldsymbol{l}|\boldsymbol{q}) = f^t_{\text{cls}}(\boldsymbol{q}) \cdot f^t_{\text{obj}}(\boldsymbol{q}), \tag{2}$$

Theoretically, given a query that represents *only* background, the objectness head should predict a very low probability of it being an object (i.e., $f^t_{\text{obj}}(\boldsymbol{q}) \approx 0$) and suppress the prediction of any objects. Conversely, if the query contains an object, then the objectness prediction should be high (i.e., $f^t_{\text{obj}}(\boldsymbol{q}) \approx 1$), and the task of classifying the query into any of the known objects or an unknown object is left to the classification head. Our challenge now becomes learning a good objectness model.

To build a robust objectness model, we turn to probabilistic models. We parametrize the objectness probability to be a multivariate Gaussian distribution in the query embedding space, i.e., $o|\boldsymbol{q} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. To predict objectness, we simply calculate the objectness likelihood, or:

$$f^t_{\text{obj}}(\boldsymbol{q}) = \exp\big(-(\boldsymbol{q}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{q}-\boldsymbol{\mu})\big) \tag{3}$$

$$= \exp\big(-d_M(\boldsymbol{q})^2\big), \tag{4}$$

where d_M denotes the Mahalanobis distance for the query embeddings. This design choice is well motivated by the current out-of-distribution (OOD) detection literature [13, 15, 26], where class-conditional Gaussian distributions are used to model the feature distribution and detected outliers. However, rather than using *class-conditional* Gaussian distributions to model the embedded feature space, we use a *class-agnostic* Gaussian distribution – as we aim to learn general object features that are shared across all classes.

Training is done in an alternating, two-step process - where we (i) estimate the distribution parameters and (ii) maximize the likelihood of matched embeddings (Fig. 2). To estimate the query embedding distribution parameters,

we estimate the batch- mean, $\boldsymbol{\mu} \in \mathbb{R}^D$, and covariance, $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ of **all** the query embeddings $\boldsymbol{Q}$ using the empirical mean and covariance estimators with exponential averaging. On the other hand, to maximize the likelihood of matched embeddings, we penalize the squared Mahalanobis distance, $d_M(\boldsymbol{q_i})^2$, of matched query embeddings, $i \in Z$. The objectness loss is, therefore, defined as:

$$\mathcal{L}_o = \sum_{i \in Z} d_M(\boldsymbol{q}_i)^2. \tag{5}$$

4.2. Objectness for Incremental Learning

In the OWOD objective, models are expected to incrementally learn newly discovered objects without catastrophically forgetting previously seen objects. To do so, OWOD methods keep a small set of images, or exemplars, to mitigate catastrophic forgetting [8, 10, 29, 31, 34]. While previous methods randomly selected instances/object classes, we believe that actively selecting instances based on their objectness score has the potential to further improve OWOD performance. Note, this does not require any extra information since we are only using existing labels, unlike in classic active learning where a model queries an oracle for additional labels. Specifically, we select instances that had either low/high objectness as exemplars. Instances with low objectness are expected to be relatively difficult instances, as the model was unsure of whether they were an object, and learning them is expected to improve the model performance on newly introduced objects. This is in line with current state-of-the-art active learning methods [28]. Meanwhile, instances with high objectness are expected to be highly representative of that object class. The selection of these instances is expected to impede catastrophic forgetting, as shown in the incremental learning field [3, 22, 33]. Specifically, after training on a particular dataset $\mathcal{D}^t$, we compute the objectness probability of every matched query embedding. We then select the top/bottom 25 scoring objects per object class. In our experiments, to avoid having an unfair advantage, if more images are selected than in previous works [8, 10], we randomly sub-sample the exemplars to match previous works.

5. Experiments & Results

We performed extensive experiments on all OWOD benchmarks, comparing PROB to all reported OWOD methods (Sec. 5.1). Extensive ablations show the importance of each one of PROB's components while shedding additional light on their function (Sec. 5.2). Finally, we test our model's incremental learning performance on the PASCAL VOC 2007 benchmark compared to other OWOD and incremental learning methods (Sec. 5.3).

Datasets. We evaluate PROB on the benchmarks introduced by Joseph *et al*. [10] and Gupta *et al*. [8], which we will reference as "superclass-mixed OWOD benchmark" (M-OWODB) and "superclass-separated OWOD benchmark" (S-OWODB) respectively. Briefly, in M-OWODB, images from MS-COCO [14], PASCAL VOC2007 [7], and PASCAL VOC2012 are grouped into four sets of non-overlapping Tasks; $\{T_1, \cdots, T_4\}$ s.t. classes in a task T_t are not introduced until t is reached. In each task T_t, an additional 20 classes are introduced - and in training for task t, only these classes are labeled, while in the test set, all the classes encountered in $\{T_\lambda : \lambda \leq t\}$ need to be detected. For the construction of S-OWODB, only the MS-COCO dataset was used and a clear separation of super-categories (e.g., animals, vehicles) was performed. However, to keep the superclass integrity, a varying number of classes is introduced per increment. For more, please refer to Joseph *et al*. [10] and Gupta *et al*. [8], respectively.

Evaluation Metrics. For known classes, mean average precision (mAP) is used. To better understand the quality of continual learning, mAP is partitioned into previously and newly introduced object classes. As common in OWOD, we use unknown object recall (U-recall), which is the ratio of detected to total labeled unknown objects [8, 19, 29, 31, 34], as mAP cannot be used (not all the unknown objects are annotated). To study unknown object confusion, we report Absolute Open-Set Error (A-OSE), the absolute number of unknown objects classified as known, and Wilderness Impact (WI). For additional details, see Gupta *et al*. [10].

Implementation Details. We use the deformable DETR [35] model utilizing multi-scale features extracted via a DINO-pretrained [2] Resnet-50 FPN backbone [8]. The deformable transformer then extracts $N_{\text{query}} = 100$ and $D = 256$ dimensional query embeddings as discussed above. The embedding probability distribution is estimated by calculating the exponential moving average of the mean and covariance of the query embeddings over the mini-batches (with a batch size of 5), with a momentum of 0.1. Additional details are provided in the appendix.

5.1. Open World Object Detection Performance

PROB's OWOD performance, compared with all other reported OWOD methods on their respective benchmarks, can be seen in Tab. 1. While all methods reported results on M-OWODB, OWOD performance on the recently introduced S-OWODB is only reported by OW-DETR. S-OWODB is expected to be a more difficult benchmark for unknown object detection, as there is complete super-category separation across the Tasks (i.e., it is more difficult to generalize from animals to vehicles than from dogs to cats). PROB shows substantial improvement in unknown object recall (U-Recall), with additional improvements in known object mAP compared to all other OWOD methods.

Table 1. **State-of-the-art comparison for OWOD on M-OWODB (top) and S-OWODB (bottom).** The comparison is shown in terms of unknown class recall (U-Recall) and known class mAP@0.5 (for previously, currently, and all known objects). For a fair comparison in the OWOD setting, we compare with the recently introduced ORE [10] not employing EBUI (EBUI relies on a held-out set of unknown images, violating the OWOD objective, as shown in [8, 34]). PROB outperforms all existing OWOD models across all tasks both in terms of U-Recall and known mAP, indicating our models improved unknown and known detection capabilities. The smaller drops in mAP between *Previously known* and *Current known* from the previous task exemplify that the exemplar selection improved our models' incremental learning performance. Note that since all 80 classes are known in Task 4, U-Recall is not computed. Only ORE and OW-DETR are compared in S-OWODB, as other methods have not reported results on this benchmark. See Sec. 5.1 for more details.

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* [10]	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	
UC-OWOD [30]	2.4	50.7	3.4	33.1	30.5	31.8	8.7	28.8	16.3	24.6	25.6	15.9	23.2	
OCPL [31]	8.26	56.6	7.65	50.6	27.5	39.1	11.9	38.7	14.7	30.7	30.7	14.4	26.7	
2B-OCD [29]	12.1	56.4	9.4	51.6	25.3	38.5	11.6	37.2	13.2	29.2	30.0	13.3	25.8	
OW-DETR [8]	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	
Ours: PROB	**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**	
ORE* [10]	1.5	61.4	3.9	56.5	26.1	40.6	3.6	38.7	23.7	33.7	33.6	26.3	31.8	
OW-DETR [8]	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	
Ours: PROB	**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**	

Table 2. **Impact of progressively integrating our contributions into the baseline.** The comparison is shown in terms of known mean average precision (mAP) and unknown recall (U-Recall) on M-OWODB. All models shown include a finetuning step to mitigate catastrophic forgetting. **PROB**−`Obj` is our model without objectness likelihood maximization. **PROB**−`L2` is our model with an L_2 loss instead of Mahalanobis distance (same as Mahalanobis distance under the assumption of $\mu = \mathbf{0}, \Sigma = \boldsymbol{I}$). **PROB**−`IL` is our model without active exemplar selection. For context, we also include the performance of deformable DETR and the upper bound as reported by Gupta *et al*. [8]. As all classes are known in Task 4, U-Recall is not computed. Additional ablations can be found in Tab. 5 of the appendix.

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	
Upper Bound	31.6	62.5	40.5	55.8	38.1	46.9	42.6	42.4	29.3	33.9	35.6	23.1	32.5	
D-DETR [35]	-	60.3	-	54.5	34.4	44.7	-	40.0	17.7	33.3	32.5	20.0	29.4	
PROB−`Obj`	**21.1**	39.3	**18.9**	41.0	23.5	32.3	**22.2**	34.7	16.3	28.6	29.2	13.4	25.2	
PROB −`L2`	22.9	53.4	19.8	49.4	28.5	39.4	21.9	37.4	15.7	30.2	30.7	14.8	26.7	
PROB−`IL`	19.4	**59.5**	15.9	54.7	**32.2**	43.5	18.4	42.6	20.7	35.3	34.7	17.4	30.4	
Final: **PROB**	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**	

Unknown Object Detection. Across all four Tasks and both benchmarks, PROB's unknown object detection capability, quantified by U-Recall, is 2-3x of those reported in previous state-of-the-art OWOD methods. This result exemplifies the utility of the proposed probabilistic objectness formulation in the OWOD objective. Other OWOD methods have attempted to integrate objectness, most notably OW-DETR [8] with their class-agnostic classification head, and 2B-OCD [29], with its localization-based objectness head (which was first reported by Kim *et al*.). As reported by Kim *et al*. [11], indeed, the utilization of their localization-based objectness estimation improves unknown object recall by $\sim$ 4-point improvement between 2B-OCD and OW-DETR. However, PROB outperformed both methods in terms of U-Recall and known mAP across all Tasks. This shows the relative robustness of our probabilistic framework compared to other methods that incorporated objectness for improved unknown object detection in the OWOD setting. When looking at Sup. Tab. 4, it becomes evident that PROB not only detects more unknowns (higher U-Recall), but it also does so much more accurately, as quantified by the reduction in A-OSE. For example, PROB had an A-OSE of 5195, 6452, and 2641 to OW-DETR's 10240, 8441, 6803 for Tasks 1-3, respectively.

Known Object Detection and Incremental Learning. PROB progressively outperforms all previous state-of-the-art OWOD methods in terms of known object mAP. Compared to OW-DETR, the method with the closest performance to ours, PROB increased known object mAP by 0.3,

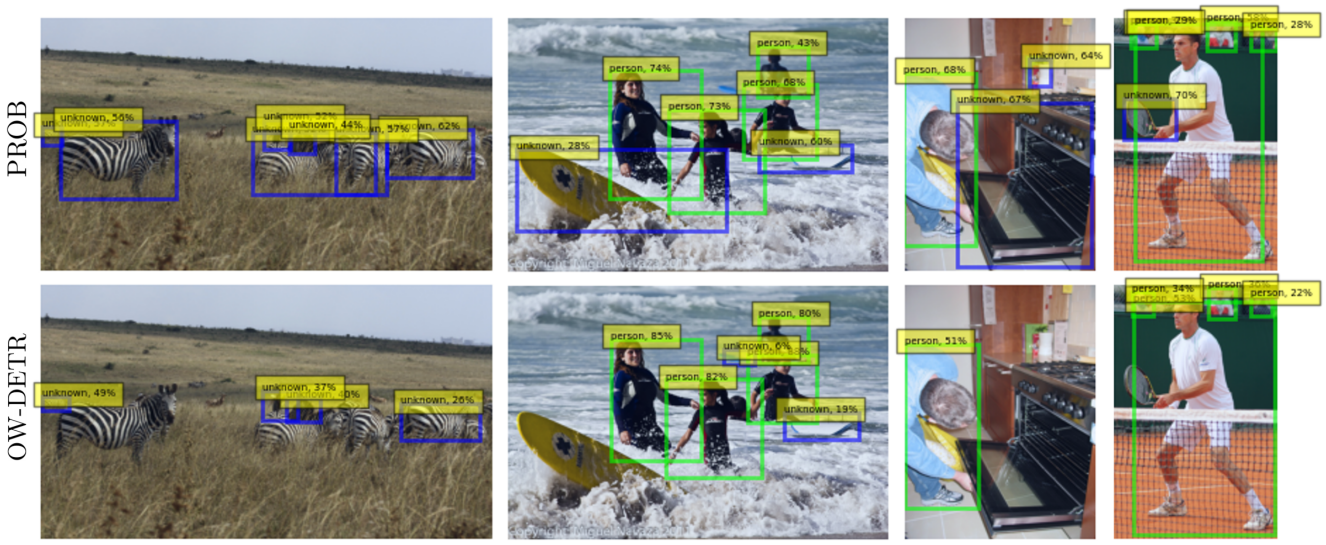

Figure 3. **Qualitative results on example images from MS-COCO test set**. Detections of PROB (top row) and OW-DETR (bottom row) are displayed, with Green - known and Blue - unknown object detections. Across all examples, PROB detected more unknown objects than OW-DETR, for example, tennis racket in the right column and zebras in the left column. Interestingly, when OW-DETR does detect unknown objects, the predictions have very low confidence, e.g., the surfing board in the center-left column.

1.1, 5.2, and 3.7 on M-OWODB Tasks 1-4 (Tab. 1, top), and 1.9, 6.6, 3.5, and 6.8 on S-OWODB Tasks 1-4 (Tab. 1, bottom). The improvement in mAP, even in Task 1, suggests that the learned objectness also improved OWOD known object handling. The relatively smaller drops between 'previously known' and the previous task's 'both' (e.g., on M-OWODB, between Tasks 1-2 OW-DETR's mAP dropped 5.6 while PROB dropped 3.8) further shows the effectiveness of PROBs active exemplar selection.

Qualitative Results. Fig. 3 shows qualitative results on example images from MS-COCO. The detections for known (green) and unknown (blue) objects are shown for PROB and OW-DETR. We observe that PROB has better unknown object performance (e.g., zebras in the left image). The unknown object predictions themselves are much more confident (oven and surfing board in the two center images). In Fig. 4, PROB detected the skateboard in Task 2 and subsequently learned it in Task 3, while OW-DETR missed both. PROB is also less prone to catastrophically forgetting an entire object class, as can be seen on the bottom of Fig. 4, where OW-DETR catastrophically forgot 'suitcase' in between Task 2 and 3, while PROB did not. These results exemplify that PROB has promising OWOD performance.

5.2. Ablation Study

Tab. 2 shows results from our ablation study. **PROB** – `Obj` disables the objectness likelihood maximization step during training, and now the objectness head only estimates the embedding probability distribution. As can be seen in Tab. 2, this had the effect of slightly increasing the unknown recall but drastically reducing the known object mAP across all Tasks. While counterintuitive, this sheds some light on how our method actually functions. Without maximizing the likelihood of the matched query embeddings, the objectness prediction becomes random. As it no longer suppresses background query embeddings, the model then predicts a lot of background patches as unknown objects and objects – known and unknown – as background. As a result, the known class mAP drops because some known object predictions are suppressed by the random objectness prediction. **PROB** – `L2` replaces the Mahalanobis distance with a standard L_2 loss, which is the same as assuming $\mu = \mathbf{0}, \Sigma = \boldsymbol{I}$. We found that using a L_2 loss resulted in a worse objectness predictor even when utilizing the same alternating optimization. Interestingly, the unknown object recall increased even when compared to **PROB** – `Obj`, showing that the model localized object bounding boxes better. As with **PROB** – `Obj`, the known class mAP drops - however not as severely. This result exemplifies the importance of the proposed probabilistic modeling approach.

PROB – `IL` disables the active exemplar selection, and it shows the advantage of using the probabilistic objectness for exemplar selection. Interestingly, it seems that the active selection mostly benefits unknown object recall, with less significant gains in both previously and currently introduced objects for Tasks 2-4. For Task 1, both methods are the same, as no exemplar replay is used. In Tab. 2, we also included the reported performance of two D-DETR models: a model trained on all classes (both known and unknown) denoted as the "Upper Bound", and a model trained and evaluated only on the known classes. Comparing the known object mAP of the oracle and D-DETR suggests that

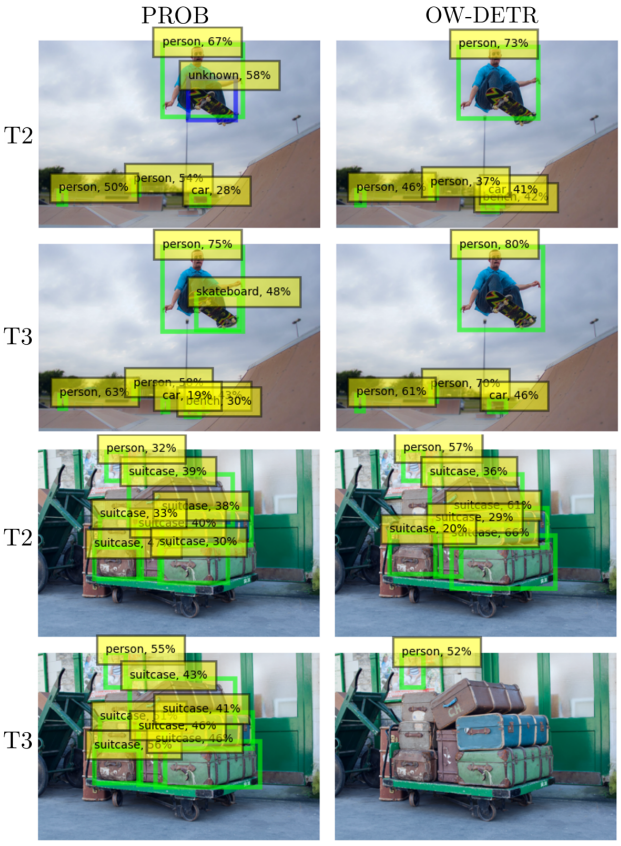

Figure 4. **Qualitative examples of forgetting and improved OWOD**. Detections of PROB (left) and OW-DETR (right) are displayed, with Green - known and Blue - unknown object detections. (top) PROB is able to detect the skateboard as unknown in T2, and subsequently classify it in T3. (bottom) Example of OW-DETR catastrophically forgetting a previously known object (suitcase).

learning about unknowns leads to improved known object detection capabilities. This possibly explains why PROB had a higher known mAP compared to OW-DETR in Task 1 on both benchmarks. Additional ablations can be found in Sec. B.3 of the appendix.

5.3. Incremental Object Detection

As reported by Gupta *et al*. [8], the detection of unknowns does seem to improve the incremental learning capabilities of object detection models. This, combined with the introduced improved exemplar selection, results in PROB performing favorably on the incremental object detection (iOD) task. Tab. 3 shows a comparison of PROB with existing methods on PASCAL VOC 2007, with evaluations performed as reported in [8, 10]. In each evaluation, the model is first trained on 10/15/19 object classes, and then an additional 10/5/1 classes are incrementally introduced. Our model had a final mAP of 66.5, 70.1, and 72.6 to OW-DETR's 65.7, 69.4, and 70.2, respectively. Results with class-breakdown are in Sup. Tab. 6 of the appendix.

Table 3. **State-of-the-art comparison for incremental object detection (iOD) on PASCAL VOC.** The comparison is shown in terms of new, old, and overall mAP. In each setting, the model is first trained on 10, 15 or 19 classes, and then the additional 10, 5, and 1 class(es) are introduced. PROB achieves favorable performance in all three settings. See Sec. 5.3 for additional details.

10 + 10 setting	old classes	new classes	final mAP
ILOD [24]	63.2	63.2	63.2
Faster ILOD [20]	69.8	54.5	62.1
ORE – EBUI [10]	60.4	68.8	64.5
OW-DETR [8]	63.5	67.9	65.7
Ours: PROB	66.0	67.2	**66.5**
15 + 5 setting	old classes	new classes	final mAP
ILOD [24]	68.3	58.4	65.8
Faster ILOD [20]	71.6	56.9	67.9
ORE – EBUI [10]	71.8	58.7	68.5
OW-DETR [8]	72.2	59.8	69.4
Ours: PROB	73.2	60.8	**70.1**
19 + 1 setting	old classes	new classe	final mAP
ILOD [24]	68.5	62.7	68.2
Faster ILOD [20]	68.9	61.1	68.5
ORE – EBUI [10]	69.4	60.1	68.8
OW-DETR [8]	70.2	62.0	70.2
Ours: PROB	73.9	48.5	**72.6**

6. Conclusions

The Open World Object Detection task is a complex and multifaceted objective, integrating aspects of generalized open-set object detection and incremental learning. For robust OWOD methods to function, understanding and detecting the unknown is critical. We proposed a novel probabilistic objectness-based approach to tackle the OWOD objective, which significantly improves this critical aspect of the benchmark. The proposed PROB integrates the introduced probabilistic objectness into the deformable DETR model, adapting it to the open-world setting. Our ablations shed light on the inner workings of our method while motivating the use of each one of its components. Extensive experiments show that PROB significantly outperforms all existing OWOD methods on all OWOD benchmarks. However, much room for improvement remains, both in unknown object detection and other aspects of the OWOD task. As such, probabilistic models may have great potential for the OWOD objective, creating powerful algorithms that can reliably operate in the open world.

Acknowledgments. OZ is generously funded by the Knight-Hennessy Scholars Foundation. We also gratefully acknowledge computational credits provided by Google Cloud Platform through Stanford's HAI Institute for Human-Centered Artificial Intelligence.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 5

[3] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 5

[4] Yinong Chen and Gennaro De Luca. Technologies supporting artificial intelligence and robotics application development. *Journal of Artificial Intelligence and Technology*, 1(1):1–8, Jan. 2021. 1

[5] Akshay Raj Dhamija, Manuel Günther, Jonathan Ventura, and Terrance E. Boult. The overlooked elephant of object detection: Open set. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1010–1019, 2020. 1

[6] 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, 2022. 1

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 5

[8] Akshita Gupta, Sanath Narayan, K J 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 (CVPR)*, pages 9235–9244, June 2022. 1, 2, 3, 5, 6, 8

[9] Mohsen Jafarzadeh, Akshay Raj Dhamija, Steve Cruz, Chunchun Li, Touqeer Ahmad, and Terrance E. Boult. Open-world learning without labels. *CoRR*, abs/2011.12906, 2020. 1

[10] K J 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 (CVPR)*, pages 5830–5840, June 2021. 1, 2, 3, 5, 6, 8

[11] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify, 2021. 2, 3, 6

[12] 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), 2021. 1

[13] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 4

[14] 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 *ECCV*, 2014. 5

[15] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020. 4

[16] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, and Laura Leal-Taixé. Opening up open world tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19045–19055, June 2022. 1

[17] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. *CoRR*, abs/1904.05160, 2019. 1

[18] Zeyu Ma, Yang Yang, Guoqing Wang, Xing Xu, Heng Tao Shen, and Mingxing Zhang. Rethinking open-world object detection in autonomous driving scenarios. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1279–1288, New York, NY, USA, 2022. Association for Computing Machinery. 1, 2

[19] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *17th European Conference on Computer Vision (ECCV)*. Springer, 2022. 2, 5

[20] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *PRL*, 2020. 8

[21] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Zhe Lin, Philip H. S. Torr, and Jiaya Jia. Open-world entity segmentation. *CoRR*, abs/2107.14228, 2021. 1

[22] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5

[23] Kuniaki Saito, Ping Hu, Trevor Darrell, and Kate Saenko. Learning to detect every thing in an open world, 2021. 3

[24] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *ICCV*, 2017. 8

[25] Deepak Kumar Singh, Shyam Nandan Rai, KJ Joseph, Rohit Saluja, Vineeth N Balasubramanian, Chetan Arora, Anbumani Subramanian, and CV Jawahar. Order: Open world object detection on road scenes. In *Proc. NeurIPS Workshops*, 2021. 1, 2

[26] Kuan-Chieh Wang, Paul Vicol, Eleni Triantafillou, and Richard Zemel. Few-shot out-of-distribution detection. *International Conference on Machine Learning (ICML) Work-*

shop on Uncertainty and Robustness in Deep Learning, 2020. 4

[27] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du Tran. Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4422–4432, June 2022. 1

[28] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9397–9406, June 2022. 5

[29] Yan Wu, Xiaowei Zhao, Yuqing Ma, Duorui Wang, and Xianglong Liu. Two-branch objectness-centric open world detection. In *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis*, HCMA '22, page 35–40, New York, NY, USA, 2022. Association for Computing Machinery. 1, 2, 3, 5, 6

[30] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. Uc-owod: Unknown-classified open world object detection, 2022. 1, 2, 6

[31] Jinan Yu, Liyan Ma, Zhenglin Li, Yan Peng, and Shaorong Xie. Open-world object detection via discriminative class prototype learning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 626–630. IEEE, 2022. 1, 2, 5, 6

[32] Shengchang Zhang, Zheng Nie, and Jindong Tan. Novel objects detection for robotics grasp planning. In *2020 10th Institute of Electrical and Electronics Engineers International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 43–48, 2020. 1

[33] Hanbin Zhao, Hui Wang, Yongjian Fu, Fei Wu, and Xi Li. Memory-efficient class-incremental learning for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5966–5977, 2022. 5

[34] Xiaowei Zhao, Xianglong Liu, Yifan Shen, Yixuan Qiao, Yuqing Ma, and Duorui Wang. Revisiting open world object detection, 2022. 1, 2, 5, 6

[35] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3, 5, 6