Decoupled PROB: Decoupled Query Initialization Tasks and Objectness-Class Learning for Open World Object Detection

Supplementary Material

Riku Inoue Masamitsu Tsuchiya Yuji Yasui
Honda R&D Co., Ltd.

{riku_inoue, masamitsu_tsuchiya, yuji_yasui}@jp.honda

1. Introduction

We report the training loss (Sec. 2), additional details of the training (Sec. 3), ablation study (Sec. 4), comparison between ETOP and DOL (Sec. 5), further discussion on the results (Sec. 6), analysis of TDQI (Sec. 7), comparison of model size and computational cost (Sec. 8), additional qualitative results (Sec. 9), and limitations (Sec. 10) in the supplementary material.

2. Training Loss

Decoupled PROB is trained with the following loss function:

$$L = L_{bb} + L_{cls} + L_{obj} \tag{1}$$

where L_{bb} denotes the L_1 and gIoU losses for bounding box learning, L_{cls} represents the sigmoid focal loss for class classification, and L_{obj} is the objectness loss(explained in main paper). All these settings are identical to those used in PROB [8].

3. Additional Training Details

We report additional training details for Decoupled PROB that are not included in the main paper. Tab. 1 lists the number of training epochs and the epochs at which the learning rate drops for each task in M-OWODB and S-OWODB. “ - ” in the learning rate drop column indicates that no special settings are applied. All other hyperparameters are the same as those used in PROB [8]. We use two Nvidia RTX A6000 GPUs for training, setting the batch size to 6 for each GPU.

4. Ablation Study

Tab. 2 shows the ablation study for Decoupled PROB. In **Decoupled PROB**-`TDQI`^{*1}, we replace Task-Decoupled Query Initialization (TDQI) in Decoupled PROB with a simple class score-based query selection [6, 7]. In **Decoupled PROB**-`TDQI`^{*2}, all object queries use only learnable parameters instead of TDQI. In **Decoupled PROB**-`ETOP`, we remove Early Termination of Objectness Prediction (ETOP) from Decoupled PROB, performing objectness prediction until the last layer of the decoder and learning without mitigating the conflict between objectness and class predictions. As shown in Tab. 2, **Decoupled PROB**-`TDQI`^{*1} exhibits lower performance across all metrics compared to Decoupled PROB, highlighting the importance of TDQI in Decoupled PROB. However, it should be noted that, similar to our proposed method, the objectness prediction is stopped at decoder layer 2(ETOP), which is the optimal setting for TDQI. Exploring the optimal number of layers for class score-based query selection could potentially improve performance. While **Decoupled PROB**-`TDQI`^{*2} improves performance for unknown objects, its performance on known objects significantly decreases compared to Decoupled PROB. As shown in Table 2 of the main paper, it can be observed that the higher the number of learnable parameters, the better the performance on unknown objects, and **Decoupled PROB**-`TDQI`^{*2} follows this trend. Similar to **Decoupled PROB**-`TDQI`^{*1}, it should be noted that **Decoupled PROB**-`TDQI`^{*2} could potentially improve performance by exploring the optimal number of layers to stop objectness prediction in ETOP. Regarding **Decoupled PROB**-`ETOP`, while it maintains comparable known object detection performance to Decoupled PROB, it shows significantly lower performance in unknown object detection. This underscores the crucial role of ETOP in enhancing the performance of Decoupled PROB.

5. Comparison between ETOP and DOL

Unlike Decoupled Objectness Learning (DOL) [2], ETOP performs both class and bounding box prediction

Table 1. **Details for training hyperparameters.**

Training Session	Task 1		Task 2		Task 2 ft		Task 3		Task 3 ft		Task 4		Task 4 ft	
	Epochs	lr drop	Epochs	lr drop	Epochs	lr drop	Epochs	lr drop	Epochs	lr drop	Epochs	lr drop	Epochs	lr drop
M-OWODB	26	15	10	-	15	10	10	-	15	10	10	-	15	10
S-OWODB	26	15	10	-	10	5	10	-	10	5	10	-	10	5

Table 2. **Comparison of ablation experiment results for the components of the proposed model.** The details of the metrics are provided in Section 5 of the main paper. In **Decoupled PROB**−TDQI^{*1}, we modify TDQI to use a purely class score-based query selection. In **Decoupled PROB**−TDQI^{*2}, we modify TDQI to use only learnable parameters. In **Decoupled PROB**−ETOP, objectness is predicted and trained until the last layer of the decoder. The performance comparison also includes Deformable DETR and the Upper Bound, which is Deformable DETR trained with ground truth for unknown classes, as reported in [1].

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 [7]	-	60.3	-	54.5	34.4	44.7	-	40.0	17.7	33.3	32.5	20.0	29.4
Decoupled PROB−TDQI^{*1}	17.9	57.4	17.5	55.1	36.0	45.6	18.8	44.1	25.5	37.9	36.9	21.2	33.0
Decoupled PROB−TDQI^{*2}	**20.9**	59.3	**18.6**	55.6	36.6	46.1	**21.0**	44.2	24.6	37.7	37.2	21.7	33.3
Decoupled PROB−ETOP	18.8	**60.4**	15.3	**56.4**	**37.2**	**46.8**	17.9	**44.7**	25.7	38.4	37.2	**22.8**	33.6
Final: **Decoupled PROB**	20.3	59.8	18.4	**56.4**	36.7	46.6	20.3	44.6	**26.1**	**38.5**	**37.8**	22.1	**33.8**

concurrently in the decoder layers that predict objectness. Additionally, ETOP incorporates iterative refinement for bounding box prediction. The combination of ETOP and TDQI offers the advantage of enabling iterative refinement across a greater number of layers through query selection.

Tab. 3 presents the experimental results comparing ETOP and DOL. TDQI+DOL indicates that ETOP in Decoupled PROB is replaced with DOL. Additionally, DOL^{*1} stops objectness prediction at the first decoder layer, while DOL^{*2} stops it at the second decoder layer. As shown in the Tab. 3, using ETOP achieves the best performance across all metrics. In TDQI, bounding box prediction is also performed during query selection, which allows iterative refinement over more layers based on those coordinates. This likely explains why ETOP demonstrates superior performance compared to DOL.

Furthermore, these results highlight the advantages of ETOP's continuous class and bounding box prediction, suggesting that it is not necessary to have layers dedicated solely to objectness prediction in Decoupled PROB.

6. Further Discussion on the Results

As shown in Table 4 of the main paper, our model often underperforms compared to CAT and USD−ASF, particularly in terms of U-Recall, on both the M-OWOD and S-OWOD benchmarks. This may be influenced by the number of object queries initialized as learnable queries. As indicated in Table 2 of the main paper, a higher ratio of learnable queries tends to improve U-Recall. Both CAT and USD−ASF initialize all object queries as learnable queries. In our approach, we use 20 object queries initialized by query selection and 80 by learnable queries.

Another reason CAT and USD−ASF may show superior metrics in early-stage tasks could be that the learnable queries in our proposed model are not fully converged yet. While our method utilizes both query selection and learnable queries, query selection is likely to focus on object surroundings from the early stages of the decoder, leading to faster convergence in training compared to learnable queries. This suggests that hyperparameter settings for our method might be more complex compared to those using only learnable queries. Although our proposed model follows the settings outlined in Tab. 1 of the supplementary material, USD−ASF and CAT undergo longer training periods (e.g., USD trains for 41 epochs and CAT for 45 epochs in task 1). Particularly in early-stage tasks such as task 1 and task 2, there is a possibility that the learnable query-based object queries in our model are not fully converged.

Additionally, one of the reasons for CAT's superior U-Recall could be that the pseudo-labeling method it uses, based on selective search, functions exceptionally well.

7. Analysis of Task-Decoupled Query Initialization (TDQI)

In TDQI, the object queries initialized with query selection are responsible for detecting known objects, while those initialized with learnable query cover missed known objects and unknown objects. We investigated the detection ratios of known and unknown classes in each task of OWODB using TDQI. The results for M-OWODB are shown in Fig. 1. As in the main paper, 20 object queries are initialized with query selection, and 80 object queries are

Table 3. **Comparison between ETOP and DOL.** DOL^{*1} and DOL^{*2} indicate that the objectness prediction in DOL stops at the first decoder layer and the second decoder layer, respectively.

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
TDQI + DOL^{*1}	19.2	59.1	17.4	53.4	33.9	43.6	19.2	43.0	23.3	36.4	36.5	20.7	32.5
TDQI + DOL^{*2}	19.8	58.2	17.0	53.4	34.0	43.7	20.0	42.2	23.7	36.0	36.0	20.6	32.1
Decoupled PROB	**20.3**	**59.8**	**18.4**	**56.4**	**36.7**	**46.6**	**20.3**	**44.6**	**26.1**	**38.5**	**37.8**	**22.1**	**33.8**

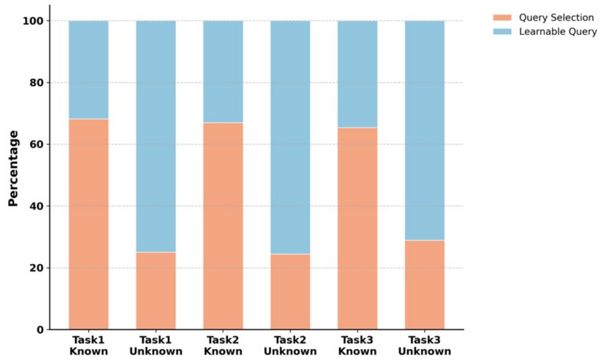

Figure 1. **Comparison of known class detection and unknown class detection roles in each task for TDQI.**

initialized with learnable query.

Although object queries initialized with query selection account for only about 20% of the total, they are responsible for nearly 70% of the detection of known classes. Conversely, learnable queries are responsible for nearly 70% of the detection of unknown objects. This demonstrates that in TDQI, object queries initialized with query selection primarily handle the detection of known objects, while those initialized with learnable query cover the missed known objects and unknown objects.

8. Comparison of Model Size and Computational Cost.

Tab. 4 presents the number of parameters and the computational cost for each OWOD model. The computational cost is calculated with an image size of 640 × 640. OW-DETR [1], CAT [4], PROB [8], and Decoupled PROB, which are based on the Deformable DETR model [7], have nearly the same number of parameters and computational costs.

OrthogonalDet [5], on the other hand, is a recently proposed high-performance OWOD model. This model has a significantly larger number of parameters and higher computational cost, and in our implementation, it outputs nearly 2000 detections (refer to the supplementary material of [5]) compared to the 100 detections output by models such as PROB and Decoupled PROB. For more details on the performance of OrthogonalDet, please refer to [5].

Table 4. **Comparison of model size and computational cost.**

Methods	**Params**	**FLOPs**
OWOD [3]	33.6M	32.1G
OW-DETR [1]	39.7M	156.3G
CAT [4]	46.1M	164.3G
PROB [8]	39.7M	156.3G
OrthogonalDet [5]	106.0M	1616.7G
Decoupled PROB (Ours)	40.9M	163.4G

9. Additional Qualitative Results

Fig. 2 illustrates examples where objects that were labeled as unknown classes in previous tasks are provided to Decoupled PROB in the current task, allowing it to learn and detect them as known classes. From left to right, the surfboard, tennis racket, sofa, and apple are detected as known classes, having transitioned from unknown classes.

Fig. 3 illustrates the qualitative results of Decoupled PROB. From left to right, it shows the reference points in the initial layer of the decoder, the reference points in the last layer of the decoder, and the detection results. The yellow reference points correspond to object queries initialized by query selection, while the green reference points correspond to object queries initialized by learnable query. The green bounding boxes indicate known classes, and the yellow bounding boxes indicate unknown classes.

10. Limitations

As shown in the results of the main paper, OWOD research has been rapidly advancing. While this task shows promise for applications in fields such as autonomous driving and robotics, there are still performance improvements needed, including in our model. These improvements include reducing false detections of background as unknown objects, improving the accurate localization of unknown objects, and preventing the forgetting of known classes.

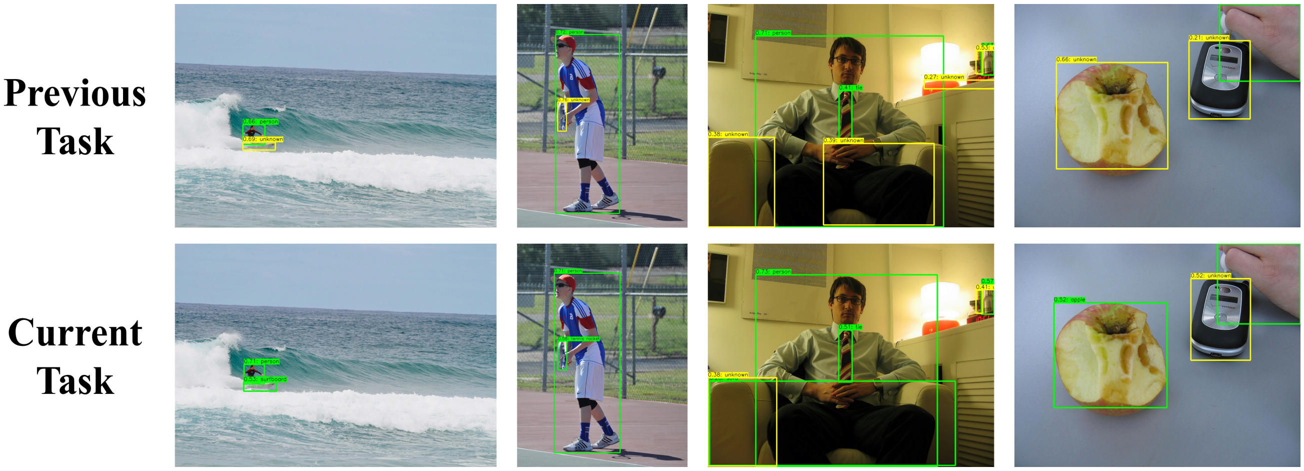

Figure 2. **Learning Unknown Classes as Known Classes (Incremental Learning).**

PROB [8], which we used as our baseline, is a remarkable approach. However, it may have an issue where background information is included in the query embeddings used to update the objectness distribution, indicating a need for future improvements. Developing new representation methods to distinguish between objects and background remains an important research challenge. Further studies are required to enable models to fundamentally understand what constitutes an object.

References

[1] 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. 2, 3

[2] Yulin He, Wei Chen, Yusong Tan, and Siqi Wang. Usd: Unknown sensitive detector empowered by decoupled objectness and segment anything model. *arXiv preprint arXiv:2306.02275*, 2023. 1

[3] 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. 3

[4] 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. 3

[5] 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. 3

[6] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021. 1

[7] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 1, 2, 3

[8] 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. 1, 3, 4

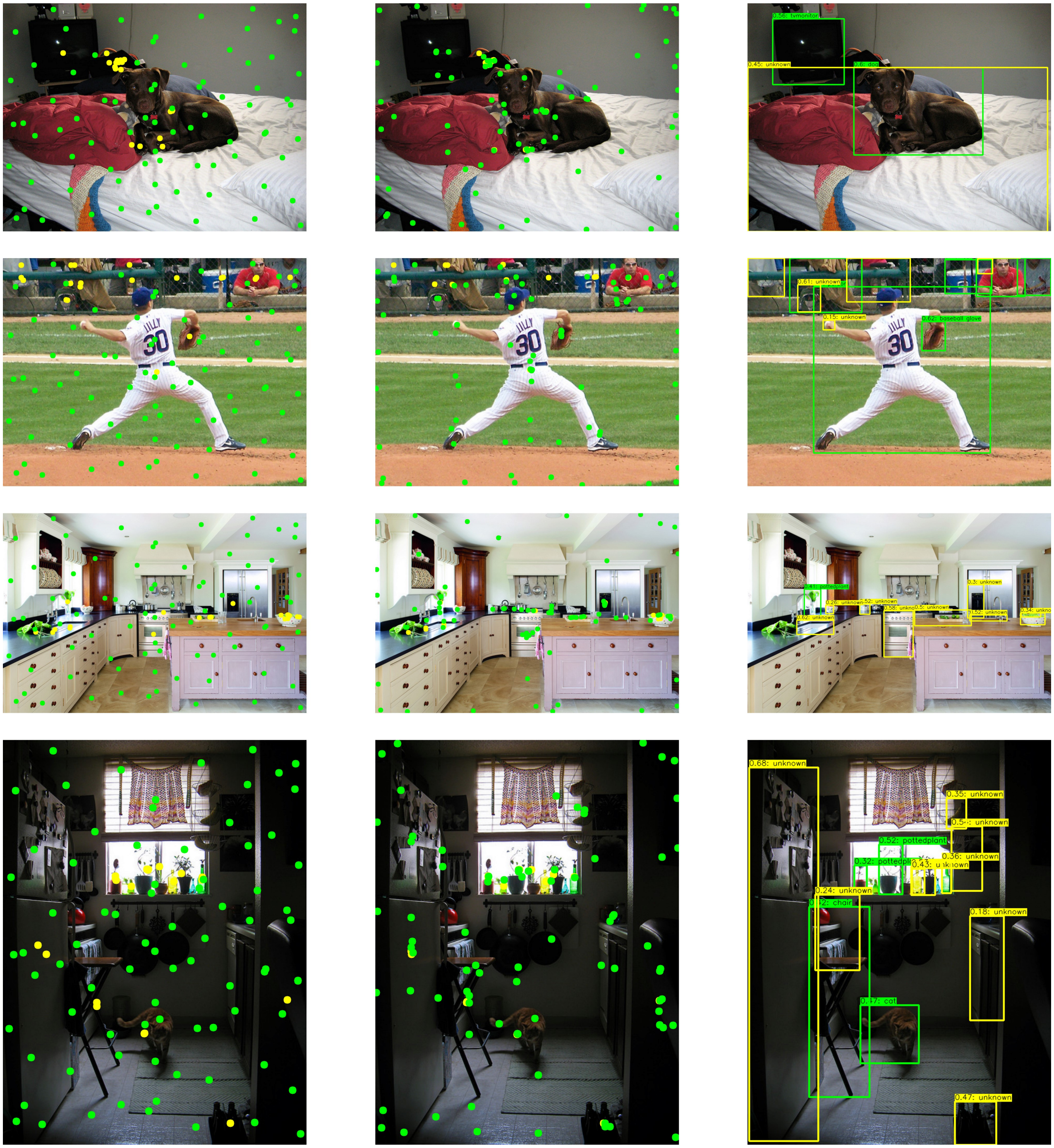

Figure 3. **Qualitative results on example images from test set.**