Supplementary Material for Annealing-based Label-Transfer Learning for Open World Object Detection

Yuqing Ma[1], Hainan Li[4], Zhange Zhang[1], Jinyang Guo[1],
Shanghang Zhang[3], Ruihao Gong[1], Xianglong Liu[1,2,4*]
[1] SKLSDE Lab, Beihang University, [2] Zhongguancun Laboratory
[3] National Key Laboratory for Multimedia Information Processing, Peking University,
[4] Institute of Data Space, Hefei Comprehensive National Science Center
{mayuqing, hainan, zhangesr, jinyangguo}@buaa.edu.cn
shanghang@pku.edu.cn, gongruihao@sensetime.com, xlliu@buaa.edu.cn

1. Analysis on Evaluation Protocols

In this part, we will elaborate on the details of different OWOD evaluation metrics and conclude that previous metrics partially evaluate the OWOD performance. Then we explain the necessity of our Equilibrium Index metric.

WI and A-OSE: The Wilderness Impact (WI) metric [2] is to evaluate the known precision variation caused by unknown misclassification:

$$\text{Wilderness Impact (WI)} = \frac{\text{P}_{\mathcal{K}}}{\text{P}_{\mathcal{K}\cup\mathcal{U}}} - 1, \tag{1}$$

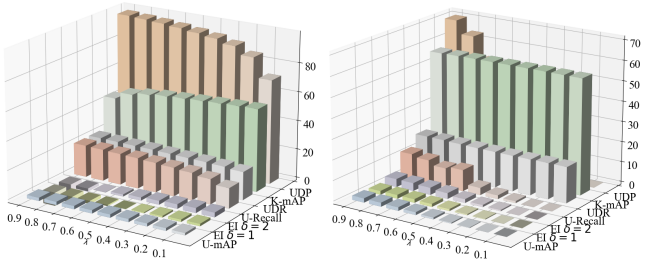
where $\text{P}_{\mathcal{K}}$ refers to the precision of the model when evaluated on known classes and $\text{P}_{\mathcal{K}\cup\mathcal{U}}$ is the precision when evaluated on known and unknown classes. The Absolute Open-Set Error (A-OSE) [6] shows the number of unknown class objects wrongly classified as known classes.

WI and A-OSE partially evaluate the OWOD performance. They measure the recognition capability from the known perspective, but neglect the known and unknown recall performance. Moreover, it cannot accurately demonstrate whether the model possesses the ability of identifying unknown objects. That is to say, even the model increases its known recognition ability with extremely low unknown detection performance, the WI and A-OSE metric are still better, which is against the objective of OWOD task.

UDR and UDP: To measure the unknown detection performance, we also adopt the evaluation metrics UDP and UDR used in [10, 11]. UDR could illustrate the localization rate of unknown objects (even misclassified as the known ones), while the UDP is the rate of correct classification of the localized unknown objects:

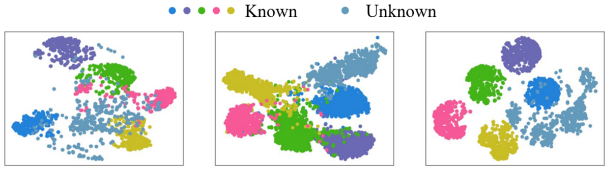
$$\text{UDR} = \frac{\text{TP}_u + \text{FN}_u^*}{\text{TP}_u + \text{FN}_u}, \tag{2}$$

*Corresponding author: Xianglong Liu



(a) Impact of fixed λ (b) Impact of different peaks of λ

Figure 1. Analysis on the disentanglement degree λ.



After forming K-mAP minimum in extending After extending

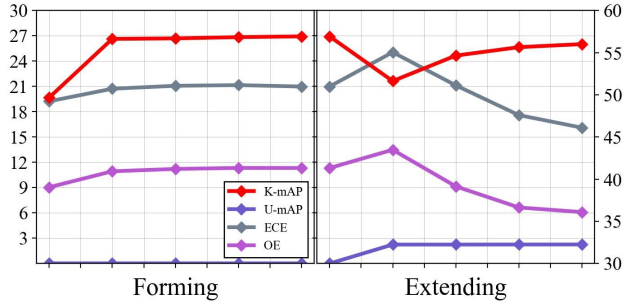
Figure 2. The t-SNE Visualization of different phases.

$$\text{UDP} = \frac{\text{TP}_u}{\text{TP}_u + \text{FN}_u^*}, \tag{3}$$

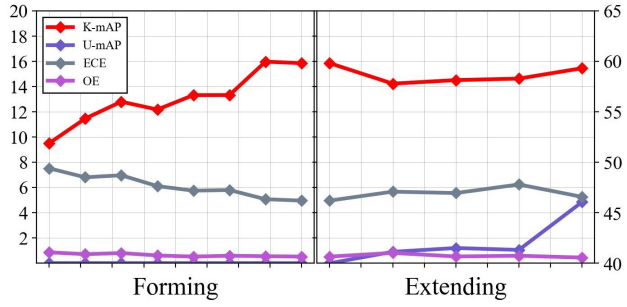
where TP_u indicates the predicted unknown boxes whose IOU with the ground-truth unknown box is more than a certain threshold (usually 0.5), and FN_u represents the missed ground-truth unknown objects. FN_u^* manifests the predicted known object whose IOU with the ground-truth unknown box is more than a certain threshold.

UDR and UDP evaluate the detection performance of unknown classes, and primarily concentrate on the objectness of the unknown categories rather than the classification precision. Nevertheless, they only consider the impacts between known and unknown classes, neglecting the influence of misclassified background.

Necessity of Equilibrium Index: From the above analysis, previous metrics partially evaluate the OWOD perfor-



(a) Ours-RCNN



(b) Ours-DETR

Figure 3. Comparison of our models with different closed world baseline models during training.

mance: WI and A-OSE only evaluate the classification performance from the known perspective, whereas UDP and UDR measure the detection performance from the unknown perspective. This conclusion is also suitable for traditional metrics, such as K-mAP (Known mAP), U-mAP (Unknown mAP), U-Recall (Unknown Recall), etc.

Therefore, it is imperative to introduce a metric that could comprehensively evaluate the OWOD performance. Moreover, we find out that in existing OWOD work, the known and unknown detection performance show a trade-off trend, which means the unknown detection gains are very close to the known performance drop.

Our Equilibrium Index simultaneously considers the known and unknown mAP variation compared to the baseline closed-world model, ensuring a fair and comprehensive comparison. We also introduce a hyper-parameter δ to adjust the concern level of unknown performance, adaptive to different scenarios.

2. Additional Experimental Analysis

2.1. Analysis on Disentanglement Degree

We further compare the performance of the model with fixed disentanglement degree λ and ours with Sawtooth Annealing Scheduling in Figure 1. The latter has already been reported in the paper, but we also put it here for easy comparison. From Figure 1 (a), we observe that the larger the fixed λ is, the better the unknown performance. On the contrary, the UDP and U-mAP will become higher as the λ increases, since the precision of the localized unknown objects is improved. Moreover, from the figure, we can observe that the U-mAP of the model with fixed λ is slightly better than ours. However, their K-mAP is less than ours. As we claimed in the paper, models with fixed λ sacrifice the recognition ability of known classes to identify the unknown ones. Although they could achieve promising unknown detection performance, they cannot guarantee the equilibrium of the known and unknown learning. In contrast, our model with annealing λ could first learn the unknown traits from transferring the known features and then collaboratively learn to identify both known and unknown objects. It validates our point that we should adopt the Sawtooth Annealing Scheduling to adjust the disentanglement degree, ensuring the known and unknown equilibrium.

2.2. Analysis on Disentanglement Process

We further analyze the disentanglement process of known and unknown classes. Figure 2 presents the t-SNE visualization of different training phases. After the forming phase, unknown features are sparsely distributed in known classes due to entanglement. At the start of the extending phase, unknown features start to disentangle from the known ones and are densely distributed. The known decision boundaries are destroyed, and the known mAP performance reaches the minimum. As the annealing scheduling, unknown and known features are co-learned. Unknown features are becoming sparse but still disentangled from known features, while known features are compact compared to those after forming. Thus, the decision boundaries are rebuilt and the known and unknown features reach equilibrium after extending. It indicates that the proposed model can accomplish the co-learning of known and unknown classes in the extending phase.

2.3. Comparison with Different Baseline Models

We implement our method based on two closed world detection models: (1) Faster RCNN [7] with ResNet-50 [4] backbone, and (2) DETR [1] following OW-DETR [3] also with ResNet-50 [4] backbone. Models with different baseline models present different detection performance. From Table 1, 2, our model with DETR yields better performance on both known and unknown classes. With DETR baseline, our model obtains nearly 5% U-mAP, significantly surpassing its counterparts. The transformer-based architecture with multi-head self-attention operation may intensify the object-level feature entanglement, activating more meaningful semantic patterns.

We also analyze the performance of intermediate models during the training. Since the unknown detection improvement may relate to the overconfidence calibration of

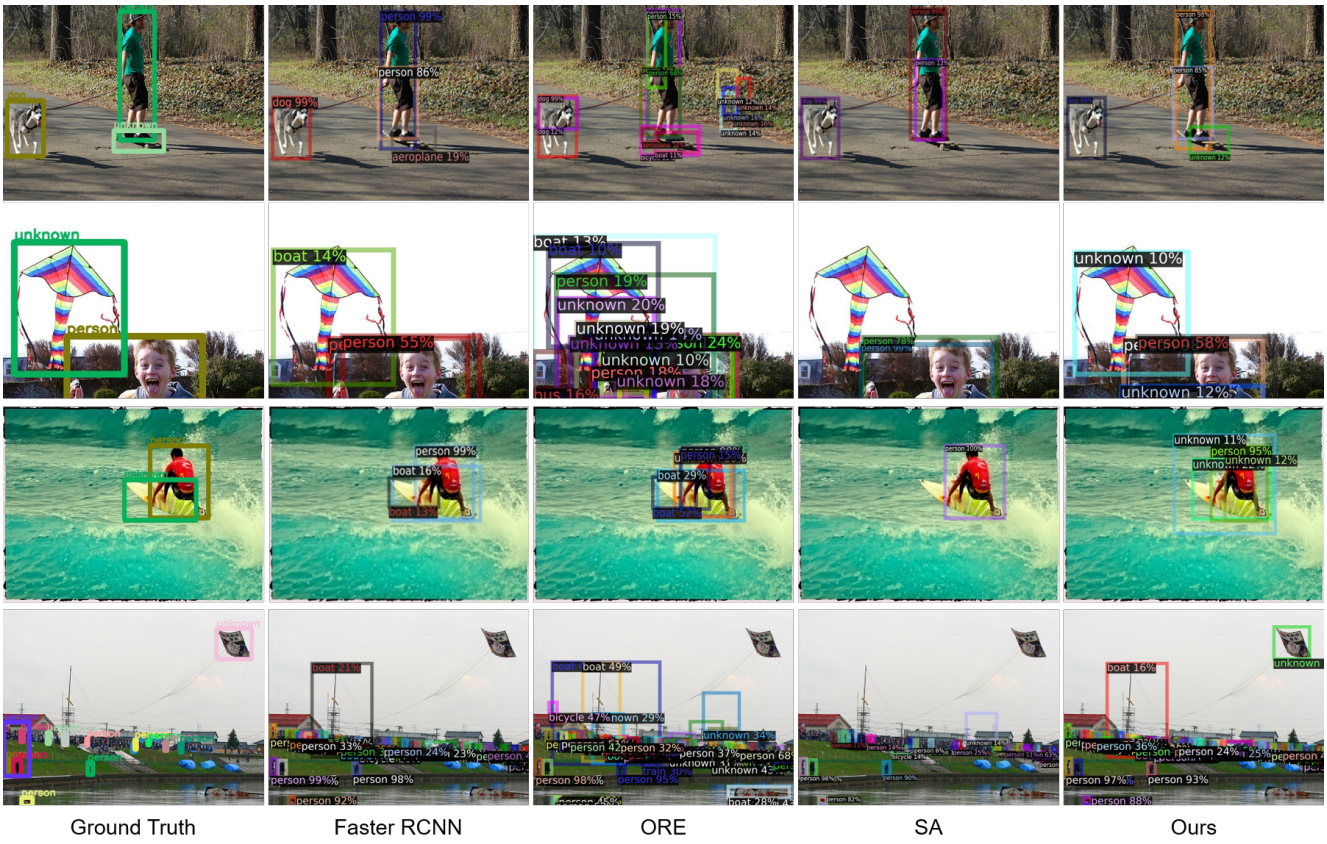


Figure 4. Visualization comparison of Faster RCNN, ORE, SA and ours.

Table 1. Comparison of our models with different closed world baseline models according to traditional detection metrics. "K-" indicates the known classes, and "U-" represents the unknown classes. The performance of closed world baselines are put at the top for reference.

Task IDs (→)	Task t_1					Task t_2					Task t_3					Task t_4
	WI-0.8 (↓)	A-OSE (↓)	K-mAP (↑)	U-mAP (↑)	U-Recall (↑)	WI-0.8 (↓)	A-OSE (↓)	K-mAP (↑)	U-mAP (↑)	U-Recall (↑)	WI-0.8 (↓)	A-OSE (↓)	K-mAP (↑)	U-mAP (↑)	U-Recall (↑)	K-mAP (↑)
Faster RCNN [7]	0.0645	10502	56.94	0	0	0.0273	8653	41.56	0	0	0.0164	7345	32.41	0	0	27.03
Ours-RCNN	0.0604	8332	56.67	2.12	12.76	0.0269	9454	40.55	0.41	5.02	0.0157	6635	32.07	0.44	9.81	27.03
DETR [12]	0.0600	57430	59.75	0	0	0.0245	27795	46.08	0	0	0.0187	17822	38.28	0	0	30.60
Ours-DETR	0.0564	46589	59.34	4.86	13.56	0.0274	24709	45.58	0.65	10.04	0.0194	14952	37.97	0.39	14.30	30.60

the known classes, we also measure the overconfidence calibration of the models according to Expected Calibration Error (ECE) and Overconfidence Error (OE) [8].

The computation of ECE and OE are listed as follows:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|, \tag{4}$$

$$\text{OE} = \sum_{m=1}^{M} \frac{|B_m|}{n} [\text{conf}(B_m) \times \max(\text{conf}(B_m) - \text{acc}(B_m), 0], \tag{5}$$

where B_m is the set of samples whose prediction scores (the winning softmax score) fall into bin m, p_i is the confidence (winning score) of the i-th sample. The $\text{acc}(B_m)$ and $\text{conf}(B_m)$ are defined the accuracy and confidence of B_m :

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \tag{6}$$

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{p}_i) \tag{7}$$

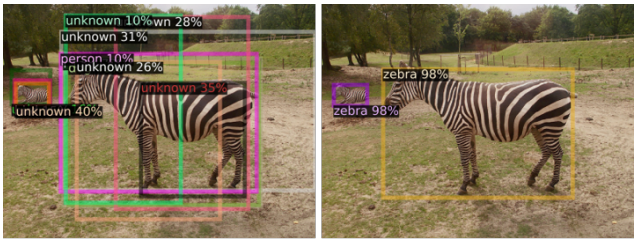
As we can see from Figure 3, in the forming phase, the RCNN-based model shows increasing known mAP and overconfidence. In contrast, although the known mAP of the DETR-based model is increasing, its overconfidence performance is declining. Then in the extending phase, with the disentangle degree λ suddenly going up to the maximum, the known accuracy drops dramatically with the ECE

Table 2. Comparison of our models with different closed world baseline models according to UDR, UDP, and our Equilibrium Index (EI).

Task IDs (→)	Task t_1					Task t_2					Task t_3				
	UDR (↑)	UDP (↑)	EI($\delta=1$) (↑)	EI($\delta=2$) (↑)	EI($\delta=5$) (↑)	UDR (↑)	UDP (↑)	EI($\delta=1$) (↑)	EI($\delta=2$) (↑)	EI($\delta=5$) (↑)	UDR (↑)	UDP (↑)	EI($\delta=1$) (↑)	EI($\delta=2$) (↑)	EI($\delta=5$) (↑)
Faster RCNN [7]	17.58	0	0	0	0	16.32	0	0	0	0	24.69	0	0	0	0
Ours-RCNN	17.95	71.08	1.85	3.97	10.33	17.62	28.49	-0.61	-0.20	1.04	23.78	41.25	0.10	0.54	1.86
DETR [12]	20.74	0	0	0	0	14.41	0	0	0	0	34.48	0	0	0	0
Ours-DETR	18.47	73.42	4.45	9.31	23.89	13.92	72.15	0.15	0.80	2.75	18.53	77.19	0.08	0.47	1.64

Table 3. State-of-the-art comparison for OWOD on incremental tasks. "Cur-" indicates the current known classes, "Prev-" indicates the previous known classes, and "U-" represents the unknown classes. We showcase models with two different closed world detection baselines, namely Faster RCNN and DETR. The performance of closed world baselines are put at the top for reference. Our model achieves superior performance in most cases.

Task IDs (→)	Task t_1			Task t_2					Task t_3					Task t_4		
	Cur-mAP (↑)	U-mAP (↑)	U-Recall (↑)	Prev-mAP (↑)	Cur-mAP (↑)	Both (↑)	U-mAP (↑)	U-Recall (↑)	Prev-mAP (↑)	Cur-mAP (↑)	Both (↑)	U-mAP (↑)	U-Recall (↑)	Prev-mAP (↑)	Cur-mAP (↑)	Both (↑)
Faster RCNN [7]	56.94	0	0	53.29	29.82	41.56	0	0	41.06	15.12	32.41	0	0	31.68	13.09	27.03
ORE [5]	56.49	0.71	5.72	**53.05**	26.22	39.64	0.14	2.66	38.67	13.16	30.17	0.12	3.34	30.16	**13.33**	25.95
SA [9]	55.56	0.20	1.93	50.31	27.73	39.02	0.03	0.79	40.38	13.86	31.54	0.003	0.12	30.99	12.72	26.42
Ours-RCNN	**56.67**	**2.12**	**12.76**	51.96	**29.13**	**40.55**	**0.41**	**5.02**	**40.82**	**14.56**	**32.07**	**0.44**	**9.81**	**31.68**	13.09	**27.03**
DETR [12]	59.75	0	0	53.78	38.37	46.08	0	0	44.01	26.83	38.28	0	0	33.54	21.76	30.60
OW-DETR [3]	58.78	0.07	7.65	52.08	36.13	44.11	0.04	5.83	41.09	25.70	35.96	0.03	5.97	32.83	13.28	27.94
Ours-DETR	**59.34**	**4.86**	**13.56**	**53.18**	**37.98**	**45.58**	**0.65**	**10.04**	**43.62**	**26.66**	**37.97**	**0.39**	**14.30**	**33.54**	**21.76**	**30.60**

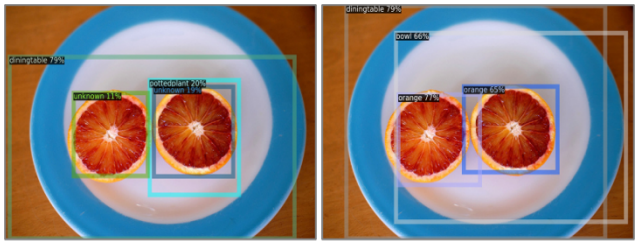


(a) (b)

Figure 5. Two zebras are identified by the model based on Faster-RCNN in Task t_1 as unknown While are correctly classified after Task t_2.

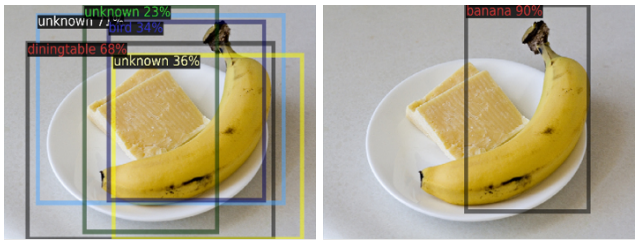
(a) (b)

Figure 6. Our model based on Faster-RCNN in Task t_1 successfully recognizes a pizza as an unknown. After learning about pizza in Task t_3, the model incrementally learns to detect it as a known.



(a) (b)

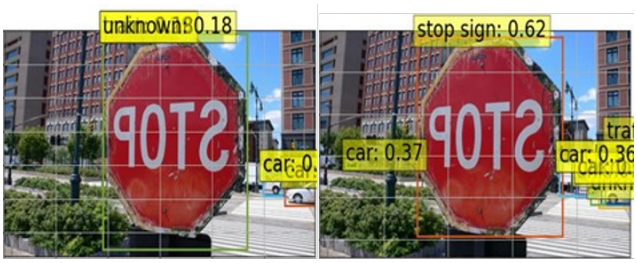
Figure 7. Orange class is not introduced in Task t_1, and the proposed model based on Faster-RCNN identifies them correctly as unknown. After learning Task t_3, these instances are labelled correctly.



(a) (b)

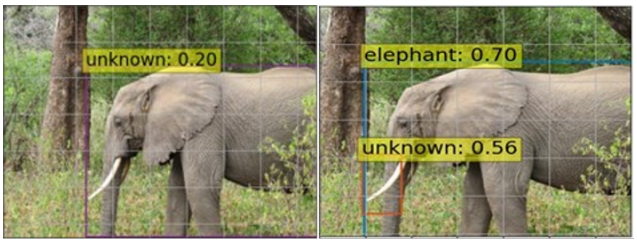
Figure 8. A banana is labeled as unknown after Task t_1. While after learning Task t_3, the model based on Faster-RCNN classifies it correctly.

and OE performance degraded. However, the DETR-based model will then return to the start point of the extending phase, while the RCNN-based model will continue to decrease to a better ECE and OE value than its initial performance of the extending phase. This phenomenon manifests that our label-transfer could calibrate the overconfi-



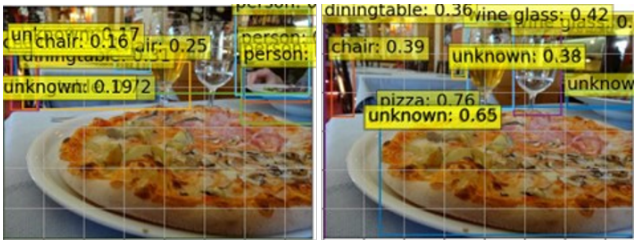
(a) (b)

Figure 9. A stop sign detected as an unknown class in Task t_1 is successfully identified in Task t_2 by our model based on DETR.



(a) (b)

Figure 10. Our model based on DETR identifies the elephant correctly as unknown as it is not introduced in Task t_1. After learning Task t_2, the instance is labelled correctly.



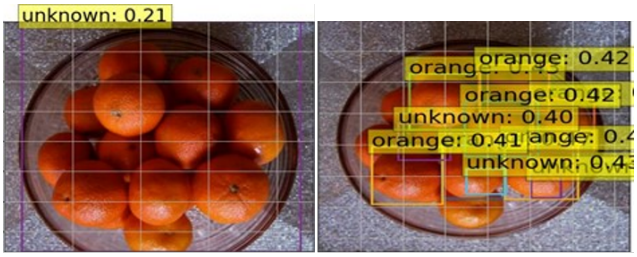
(a) (b)

Figure 11. A pizza is recognized as unknown after Task t_1 and is classified correctly after learning Task t_3.

dence of the RCNN-based model to a certain extent, and the DETR-based model, even without our label-transfer learning, shows better overconfidence calibration ability. Moreover, consistent with our intuition, the unknown performance increases and stays stable with the disentanglement degree λ annealing. Even the λ is diminishing, as long as it exists, the model would simultaneously learn both known and unknown traits, and dynamically reach the equilibrium.

2.4. Ablation Study based on DETR

We investigate the effectiveness of different components in the proposed framework based on DETR architecture. Table 4 respectively lists the performance of our model



(a) (b)

Figure 12. Some oranges are labeled as unknown after Task t_1. While after learning Task t_3, the proposed model based on DETR classifies them correctly.

Figure 13. Failure cases of our model.

Table 4. Ablation study of Our model based on DETR. The full model (SAS+LT) yields the superior performance, and each module contributes to the proposed model.

	EI($\delta = 1$)	K-mAP	U-mAP	UDR	UDP
w/o LT	0	59.75	0	20.74	0
w/o SAS	0.54	58.17	2.12	18.93	89.76
full FT	-50.38	0	9.37	19.97	89.52
with NC	-2.92	56.24	0.59	21.08	99.96
Ours(SAS+LT)	4.45	59.34	4.86	18.47	73.42

without Label-Transfer Learning (w/o LT) which degrades to the DETR model, our model without Sawtooth Annealing Scheduling (w/o SAS), our model with full Label-Transfer (full LT) where all known proposals will be projected into unknown classes, our model with a manual unknown selection strategy Novelty Classification proposed by OW-DETR [3] (with NC), and the full model (SAS+LT) based on DETR. As we can see from the table, the DETR-based model without the Label-Tansfer lacks the ability to detect unknown classes. Without SAS, the detection ability of

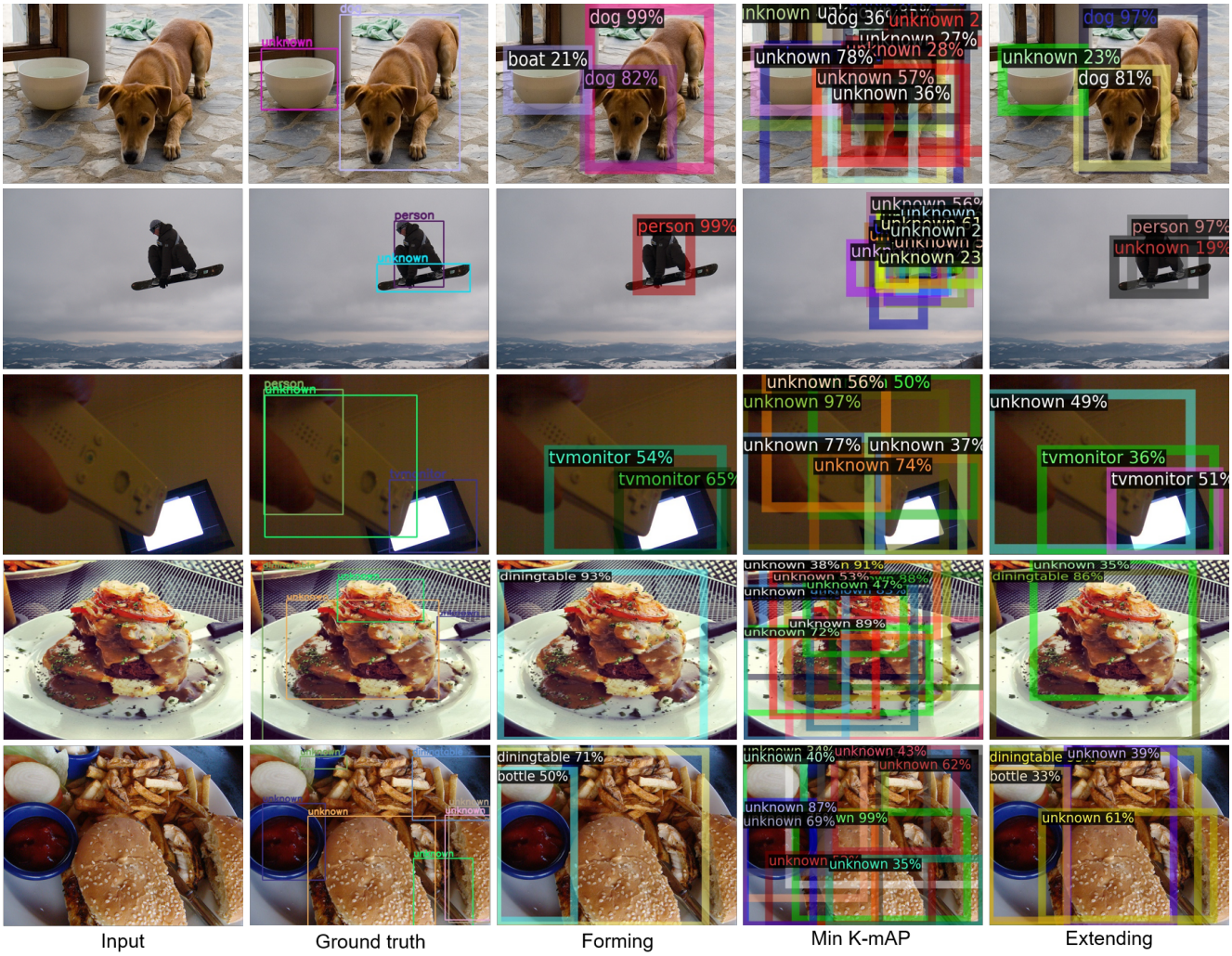


Figure 14. Visualization of different training phases.

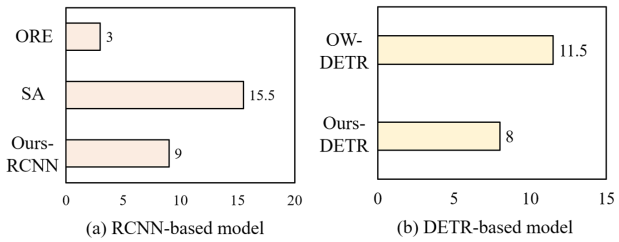
the model for both known and unknown classes decreases, due to the lack of annealing process which is designed to achieve collaborative learning. Full Label-Transfer will greatly damage the known detection performance and even causes the DETR-based model to lose the detection capability of known classes. Besides, adopting the unknown discovery strategy NC cannot bring performance gains for our model. In conclusion, our full model based on DETR achieves the best performance, and each module contributes to the proposed model.

2.5. Analysis of Time Complexity

Figure 15 illustrates the training time in task t_1 of the state-of-the-art OWOD methods and ours. From the table, we can observe that, for models based on the RCNN structure, ORE with a simple unknown-selection strategy costs nearly 3.0 hours, while ours costs 9.0 hours but less than the newly-proposed method SA (15.5 hours). For models based on the DETR structure, the training time of our model (8.0 hours) is less than OW-DETR (11.5 hours). That is because our model does not require manually selecting unknown proposals, while OW-DETR proposed a complex attention-driven unknown-selection strategy with more computations. Considering our detection performance gains, our training time is totally acceptable, and even more efficient compared to the newly-proposed OW-DETR and SA.

2.6. Analysis on Incremental Tasks

Due to the space limitation, we only report the known performance of both previous known classes and the current known classes (namely, the newly-annotated classes). In Table 3, we list the detailed information on known detection performance. In most cases, our model achieves superior performance, proving that our model could main-



(a) RCNN-based model (b) DETR-based model

Figure 15. State-of-the-art comparison of training time (hours) in Task t_1.

tain the known performance with unknown recognition improvement.

3. Additional Visualization

Figure 4 shows the visualization comparison with other state-of-the-art models based on the Faster RCNN baseline. We can observe that our model could simultaneously detect both known and unknown objects with more accurate bounding boxes and classification labels. In addition, with the unknown identification ability, ORE tends to overgenerate bounding boxes aiming at localizing more unseen objects. SA attempts to under-generate bounding boxes to avoid misclassification. In contrast, without manually-designed unknown-discovery strategies, our model could produce appropriate bounding boxes and make precise predictions.

Figure 5 to Figure 12 list more visualization comparisons of our model in different tasks. Figure 5 to Figure 8 show the results of the RCNN-based model, and Figures 9 to 12 show the results of the DETR-based model. Some objects are detected as unknown classes in Task t_1 and will be recognized in the following tasks. Figure 13 further displays some failure cases of our model. Without appropriate guidance, the model easily classifies part of the known objects or multiple unknown objects as a single unknown instance. Moreover, with nearly 20% UDR performance, there must be many missing-detection situations.

We also exhibit visualization of each phase during training in Figure 14. At the end of the forming phase, only objects of known classes could be identified. In the middle of the extending phase when the known mAP reaches its lowest point, most objects are recognized as the unknown class. Finally, after the disentangle degree λ returns to zero, the model is thoroughly trained, and both known and unknown objects can be detected.

4. Supplementary Video

We further show the disentanglement process and more visualization of our approach in the supplementary video, which is available at the google drive link.

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

[2] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1021–1030, 2020. 1

[3] 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, 4, 5

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] 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. 4

[6] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3243–3249. IEEE, 2018. 1

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 3, 4

[8] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[9] Shuo Yang, Peize Sun, Yi Jiang, Xiaobo Xia, Ruiheng Zhang, Zehuan Yuan, Changhu Wang, Ping Luo, and Min Xu. Objects in semantic topology. *arXiv preprint arXiv:2110.02687*, 2021. 4

[10] Xiaowei Zhao, Xianglong Liu, Yifan Shen, Yuqing Ma, Yixuan Qiao, and Duorui Wang. Revisiting open world object detection. *arXiv preprint arXiv:2201.00471*, 2022. 1

[11] Jiyang Zheng, Weihao Li, Jie Hong, Lars Petersson, and Nick Barnes. Towards open-set object detection and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2022. 1

[12] 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. 3, 4