Taming Self-Training for Open-Vocabulary Object Detection

Supplementary Material

This document is supplementary to the main paper as,

- Sect. A.1 elaborates how to train a detector for open-vocabulary object detection (OVD) without an external RPN, which is used in "**External RPN**" of Sect. 4.3 in the main paper.
- Sect. A.2 shows the quality of pseudo labels (PLs) from the initial teacher and how to improve initial PLs for Sect. 3.1.
- Sect. A.3 elaborates the exact iterations when we update the teacher model for "**The number of updates to the teacher model.**" of Sect. 4.3.
- Sect. A.4 shows the performance of our method with different initial weights.
- Sect. B.1 compares the proposed SAS-Det and visual grounding methods.
- Sect. B.2 shows how well SAS-Det preserves knowledge learned in the pretraining.
- Sect. B.3 provides extra ablation studies on LVIS.
- Sect. B.4 explores stronger external RPNs for SAS-Det.
- Sect. B.5 explores to use external PLs at the beginning of self-training
- Sect. C.1 visualizes good and failure cases of PLs.
- Sect. C.2 visualizes good and failure cases of our OVD detector.
- Sect. D discusses some limitations of this work and potential solutions.

A. Extra Details for The Main Paper

A.1. Our detector for OVD without an external RPN

This section explains how we train a detector without an external RPN. This approach is compared with the *baseline* in paragraph "**External RPN**" of Sect. 4.3. As shown in Fig. 1, we divide the training into two stages: (a) In the first stage, we put a RPN box head on top of a pretrained but frozen CLIP image encoder and only train the box head with a box loss. The RPN box head outputs region proposals with objectness scores. The box loss consists of a foreground classification term and a box regression term. (b) In the second stage, the RPN box head and a detection head are built on the same CLIP image encoder. No modules are frozen. The whole model is trained the same way as *baseline* in the main paper with online pseudo labels from a teacher model.

Table 1 provides the performance of the first and the second training stages. We evaluate the model of the 1st stage by directly classifying region proposals with text embeddings. The model of the 1st stage can be regarded as the initial teacher of the 2nd stage. It achieves similar performance as the initial teacher of *baseline*, which indicates that initial PLs of *baseline* and the 2nd stage share similar qualities. However, *baseline* outperforms the 2nd stage's model on both base and novel categories. Such results clearly demonstrate that an external RPN is important to our detector and training.

A.2. Improving initial pseudo labels with RPN scores

In the self-training process, we initialize the teacher model with pretrained CLIP weights to generate PLs. However, the initial teacher cannot provide high quality PLs because CLIP is weak at localizing objects and has poor zero-shot detection ability [5, 14]. Similar to VL-PLM [14], we average the objectness scores from the external RPN with the prediction scores from the teacher model. Assuming s_i^{RPN} is the objectness score of the i-th region proposal, the averaged prediction score is $\hat{p}_{i,c} = (s_i^{\mathrm{RPN}} + p_{i,c})/2$ where c refers to the c-th category. Table 2 provides the quantitative results for whether or not to use the RPN fusion. As shown, without the fusion, the quality of PLs significantly drops.

A.3. When to update the teacher model.

In this section, we discuss the timing of updates to the teacher model during training. We trained our detectors with different times of updates to teacher models and provides the exact iterations when we update the teacher model for COCO-OVD in Table 3. Generally, we consider the learning rate schedule and distribute updates as evenly as possible during training. As shown in Table 3, too many updates, e.g., 8 or 4 updates, lead to performance drops mainly due to the following. First, similar as the aforementioned EMA update, too many updates change the distribution of PLs too often and make the training unstable. Second, the more updates, the earlier an update happens. However, the student model is not well trained at the early stage of the training and thus is not good enough to update the teacher. Table 3 shows that 2 and 3 updates achieve similar performance. But we set 3 updates as default to include as many updates as possible, considering that the only overhead of our update is to copy the weights from the student to the teacher. For LVIS-OVD, we find that a later update helps and only conduct the update when the learning rate changes. For the 2× training setup, the teacher is updated at 120k and 160k iterations.

A.4. Initial weights.

We use CLIP weights from RegionCLIP [15] to initialize SAS-Det. However, we noticed that RegionCLIP's pretrain-

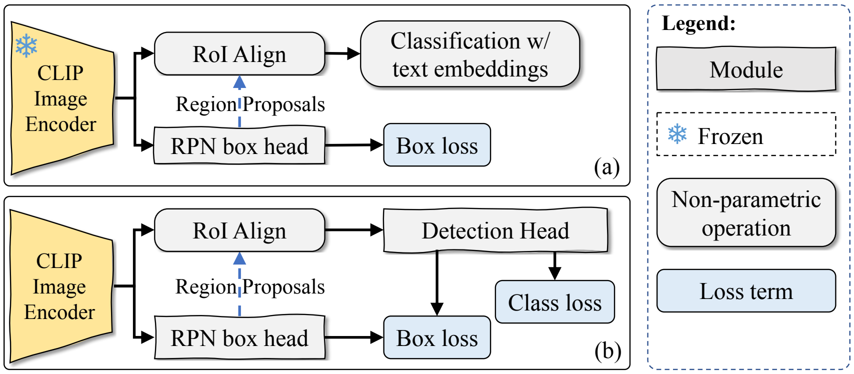


Figure 1. Two stage training for an OVD detector without an external RPN. **(a)** In the first stage, only RPN box head is trained. Text embeddings are used for classification at inference time. **(b)** In the second stage, no modules are frozen.

Models	AP$_{50}^{novel}$	AP$_{50}^{base}$	AP$_{50}^{all}$
First stage (no external RPN)	10.5	12.7	12.1
Second stage (no external RPN)	25.4	53.4	46.1
Initial teacher of *baseline* (w/ an external RPN)	11.9	-	-
baseline (w/ an external RPN)	**31.4**	55.7	49.4

Table 1. Performance of OVD detectors without an external RPN.

Method	AP$_{50}^{novel}$ (COCO)	AP$_r$ (LVIS)
w/o RPN scores	3.8	1.9
w/ RPN scores	**11.9**	**9.1**

Table 2. The quality of PLs generated by the initial model w/ or w/o the RPN fusion on novel categories of the COCO and LVIS datasets.

# Updates	Iterations to update	AP$_{50}^{novel}$
8	Every 1k iterations	27.6
4	Every 2k iterations	30.6
3 (*baseline*)	4k,6k,8k	31.4
2	4k,8k	**31.6**
1	5k	30.9
0 (No update)	N/A	29.6

Table 3. Iterations to update the teacher model for different number of updates on COCO-OVD. The updates are usually conducted at 6k and 8k iterations when the learning rate decreases.

ing includes LVIS base boxes that may include boxes for novel categories of COCO. To avoid potential data leakage, for our experiments on COCO-OVD, we followed RegionCLIP's procedure to finetune CLIP with COCO base boxes only. Besides, On COCO-OVD, we report results with Soft NMS [1] as RegionCLIP [15] did.

Table 4 compares original CLIP weights with RegionCLIP weights as the initialization of *baseline*. As shown, initialized with either of them, the detectors achieve similar performance on novel categories on both the COCO and LVIS datasets. This is probably because our finetuning with PLs closes the gap between CLIP's and RegionCLIP's pretraining. Since RegionCLIP does not benefit our method, it is still fair to compare SAS-Det with other methods that uses the original CLIP. Table 4 also shows that initialization with RegionCLIP improves the performance on base categories. Since RegionCLIP adopts the boxes of base categories in its pretraining, it actually provides longer training on base categories. We attribute the improvement on base categories to the longer training.

B. Additional Experiments

B.1. Comparison with visual grounding models

There are some recent studies [2, 7, 9] focusing on large-scale pretraining for visual grounding where text phrases in a whole caption are aligned with objects. They usually train their models with multiple datasets, including detection data, visual grounding data [9], and image-text pairs. Visual grounding data requires the association between each box and each specific text phrase, which is much more expensive than detection annotations. The datasets GoldG+ and GoldG, introduced in [9],are two widely used Visual grounding datasets. GoldG+ contains more than 0.8M human-annotated gold grounding data curated by MDETR [7], including Flickr30K [10], VG Caption [8], GQA [6], and RefCOCO [13]. GoldG removes RefCOCO from GoldG+. Image-text pairs are usually from CC [12] and LocNar [11].

Table 5 compares our method with MDETR [7], GLIP [9] and X-DETR [2] on LVIS minival that contains 5k images. As shown, our detector with RN50x4 as the backbone achieves the leading performance on rare cate-

Initial weights	COCO-OVD			LVIS-OVD			
	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}^{all}	AP_r	AP_c	AP_f	AP
From original CLIP	31.1	53.5	47.7	19.6	23.6	30.6	25.6
From RegionCLIP pretraining	**31.4**	55.7	49.4	**19.8**	25.3	31.5	26.8

Table 4. Performance of *baseline* using different pretrained weights as initialization.

Method	Training Data	Backbone	Detector	AP_r	AP_c	AP_f	AP
MDETR	LVIS, GoldG+	RN101	DETR [3]	7.4	22.7	25.0	22.5
GLIP	O365, GoldG, Cap4M	Swin-T	DyHead [4]	20.8	21.4	31.0	26.0
X-DETR	LVIS, GoldG+, CC, LocNar	RN101	Def.DETR [16]	<u>24.7</u>	34.6	35.1	34.0
Ours	LVIS Base	RN50-C4	FasterRCNN	20.9	26.1	31.6	27.4
Ours	LVIS Base	RN50x4-C4	FasterRCNN	**27.3**	30.7	35.0	31.8

Table 5. Comparison with visual grounding methods on LVIS minival. Our method adopts pretrained CLIP model that is supervised with image-text pairs automatically collected from the Internet. Reference for methods: MDETR [7], GLIP [9], X-DETR [2]

gories without using any annotations of those categories. By contrast, the other three methods employ the costly visual grounding data. MDETR [7] and X-DETR [2] even adopt the ground truth of rare categories but cannot outperform our method. Such results further demonstrate the effectivenss of the proposed SAS-Det.

B.2. Preserving the knowledge from the pretraining

In this section, we explore if models after our finetuning generalizes as well as pretrained models. If so, our finetuning preserves the knowledge learned in the pretraining. Specifically, we evaluated the proposed SAS-Det and *baseline*, which are trained on COCO-OVD with 65 concepts, on LVIS with 1203 concepts. Then, we compare them with the pretrained models and report the results in Table 6. As shown, though finetuned with limited concepts, SAS-Det and *baseline* achieve similar or better performance as pretrained models on rare categories of LVIS. Note that those categories do not appear during finetuning. This indicates that the knowledge learned in the pretraining is successfully preserved in our finetuning. We attribute the improvement on common and frequent categories to the fact that finetuning adapts CLIP to OVD and thus the models learn how to better handle instance-level detection instead of image-level classification that CLIP is pretrained for.

B.3. Ablation studies on LVIS

In this section, we provide ablation studies on LVIS, which are similar as what we did on COCO. Except trained on LVIS-OVD, the baseline model *LVIS baseline* is the same as *baseline* in the main paper. As shown in Table 7, we have consistent observations on LVIS as on COCO. First, based on the results of *LVIS baseline* and *(1)*'s, it is beneficial to remove noisy pseudo boxes from finetuning. Second, *(3)*'s outperforms both *LVIS baseline* and *(2)*'s, which demonstrates that the proposed SAF head helps. The improvement probably comes from the fact that the SAF head avoids noisy pseudo boxes as supervision and incorporates a fusion from different branches.

Method	AP_r	AP_c	AP_f	AP
Original CLIP	8.7	6.7	4.0	6.0
RegionCLIP pretraining	9.7	7.1	4.3	6.4
baseline	9.2	9.7	9.0	9.4
SAS-Det (Ours)	**13.1**	**10.5**	**9.9**	**10.7**

Table 6. Preserving the knowledge from the pretraining. Models are trained on COCO-OVD and evaluated on the LVIS validation set. Most LVIS categories are unseen during training.

B.4. Inference with different RPNs

The proposed SAS-Det leverages an external RPN to get region proposals, and thus it is open to other RPNs without any further finetuning. As shown in Table 8, our model is evaluated together with several RPNs that are trained with different data. Based on those results, we have several findings. First, the model is only trained with *(4)*'s RPN but achieve similar or better performance with other RPNs. This indicates that our model does not rely on the specific RPN that is used in the training, and it is open to different RPNs. Second, *(4)*'s performance is close to others on novel categories, but its RPN is trained without any boxes of novel categories. This demonstrates that the RPN trained on base categories generalizes to novel ones.

B.5. Using external PLs at the beginning

At the beginning of self-training, our PLs are not as good as VL-PLM's [14], but our training pipeline allows us to leverage external high quality PLs before the first update

Ablation	AP_r	AP_c	AP_f	AP
LVIS baseline	19.8	25.3	31.5	26.8
(1) Use noisy pseudo boxes to train box regression	(-4.5) 15.3	24.0	31.1	25.2
(2) No pseudo labels, train with base data only	(-3.1) 16.7	27.0	33.0	27.6
(3) SAF head, fuse the open- and the closed-branches	(+1.1) 20.9	26.1	31.6	27.4

Table 7. Ablation studies to analyze the effect of components of SAS-Det on LVIS-OVD.

Training boxes for RPNs	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}^{all}
(4) COCO Base (48 categories)	31.4	55.7	49.4
(5) COCO Base + Novel (65 categories)	32.7	55.7	49.7
(6) COCO (80 categories)	32.9	55.7	49.8
(7) LVIS Base (866 categories)	32.9	55.2	49.3

Table 8. Evaluations with different RPNs on COCO-OVD.

Method	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}^{all}
baseline	31.4	55.7	49.4
baseline + VL-PLM's PLs	33.5	55.9	50.1

Table 9. Using PLs of VL-PLM [14] in our self-training. Results on COCO-OVD are reported.

to the teacher. To demonstrate this, we trained *baseline* on COCO-OVD using VL-PLM's PLs before the first update. As shown in Table 9, with VL-PLM's PLs, our model improves from 31.4 to 33.5 for novel categories.

C. Qualitative Results

C.1. Visualizations of pseudo labels

Figure 2 provides failure cases of our final PLs (after all updates) on the COCO dataset. We find two major types of failures. (a) Redundant boxes. In this case, one object has multiple predictions that are overlapped with each other. Those overlapped PLs indicate that the pseudo boxes are extremely noisy and cannot be improved by a simple thresholding based on classification confidences. Thus, it is necessary to handle the noise in pseudo boxes separately. We believe that redundant PLs are caused by the poor localization ability of CLIP that provides noisy initial PLs. Though our finetuning improve the localization ability to some extent, how to further improve this ability is still challenging and requires future research. (b) Wrong categories. We find the teacher model tends to classify every object into given concepts and generates PLs with wrong categories. Fortunately, there are usually some connections or similarities between the detected objects and the wrong categories. OVD employs text embeddings as classifiers, and text embeddings of related concepts share similarities. For example, the embeddings of "bus" and "train" are close to each other. Thus, though with wrong categories, those PLs may still provide supervision for OVD to some extend.

Figure 3 visualizes more PLs with different times of updates to the teacher model. All samples come from the COCO dataset. As shown, PLs before the update are noisy. Updates remove the noise and improve the quality of PLs.

C.2. Visualizations of our detector for OVD

We visualize good cases and failure cases of the final detector in Fig. 4 and Fig. 5, respectively. All samples come from the COCO dataset, and only predictions of novel categories are provided. As shown in Fig. 4, our detector is able to detect rare objects, e.g. a toy umbrella, a bus with rich textures, and an elephant sculpture.

We find two major failure cases as shown in Fig. 5. (a) Missing instances that usually happen in images with a crowd of objects belonging to one category. In our view, such cases are difficult for fully supervised object detection, let alone OVD. (b) Redundant predictions. We believe this is caused by using redundant PLs (see examples in Fig. 2a) as supervision. Improving PLs will alleviate redundancy in predictions.

D. Limitations

This work has the following limitations that can be further investigated: (1) Compared to standard training, our self-training involves an additional teacher model, which requires more GPU memory. One possible solution to reduce the cost is to alternatively run the teacher and the student. (2) Although much faster than prior methods, our online pseudo labeling still induces overhead during training when millions of iterations are required. One possible solution is to generate offline PLs each time the teacher model is updated. (3) Although achieving good performance, SAS-Det still suffers from two major failure cases, which are visualized in Sect. C.2. The future work may explore stronger pretrained VLMs and better denoising steps for solutions.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-NMS–improving object detection with one line of code. In *ICCV*, pages 5561–5569, 2017. 2

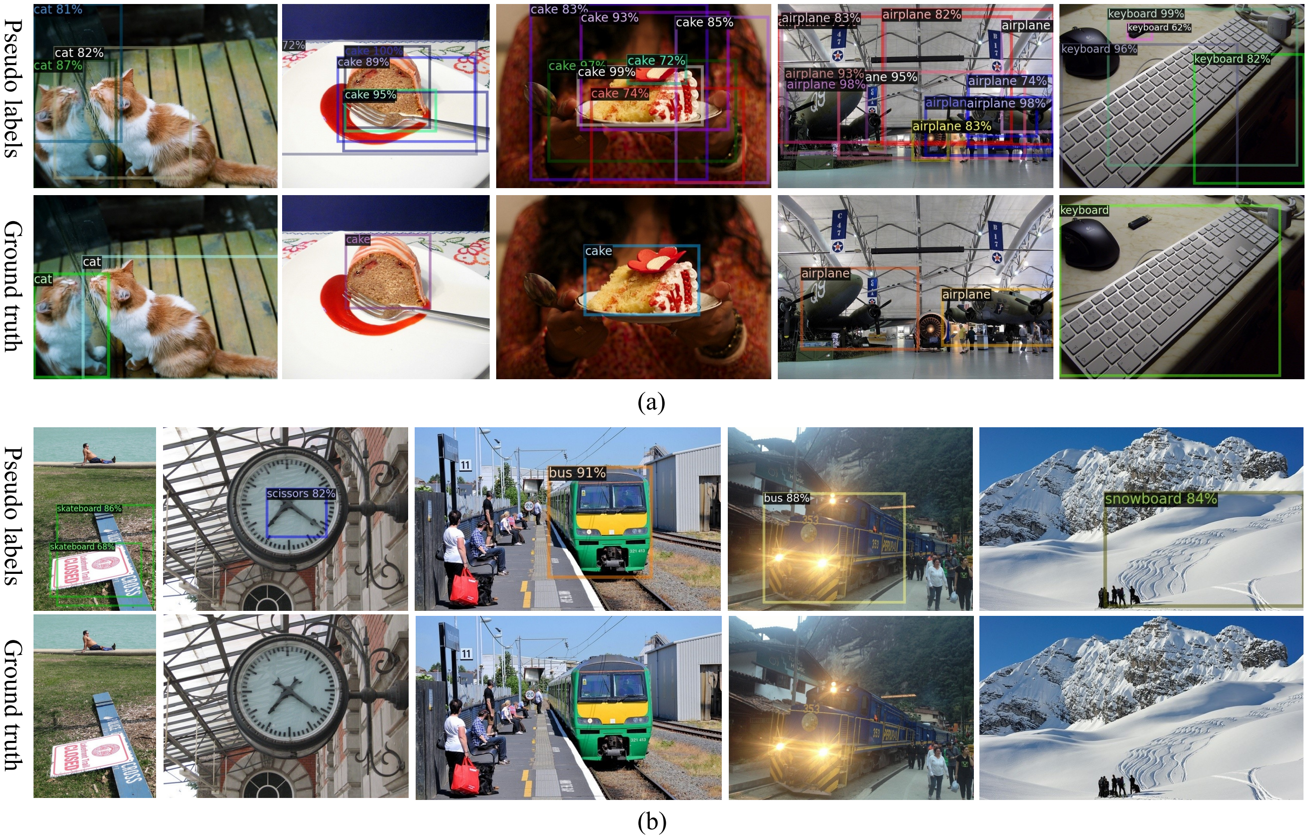


Figure 2. Visualizations of failure cases in PLs after three updates. All samples are from COCO. Two major types of failures: **(a)** Redundant boxes. **(b)** Wrong categories.

[2] Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-DETR: A versatile architecture for instance-wise vision-language tasks. In *ECCV*, pages 290–308. Springer, 2022. 2, 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *ECCV*, 2020. 3

[4] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *CVPR*, pages 7373–7382, 2021. 3

[5] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*, 2022. 1

[6] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 2

[7] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR – Modulated Detection for End-to-End Multi-Modal Understanding. In *ICCV*, 2021. 2, 3

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 2

[9] 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 *CVPR*, pages 10965–10975, 2022. 2, 3

[10] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 2

[11] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, pages 647–664. Springer, 2020. 2

[12] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 2

[13] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016. 2

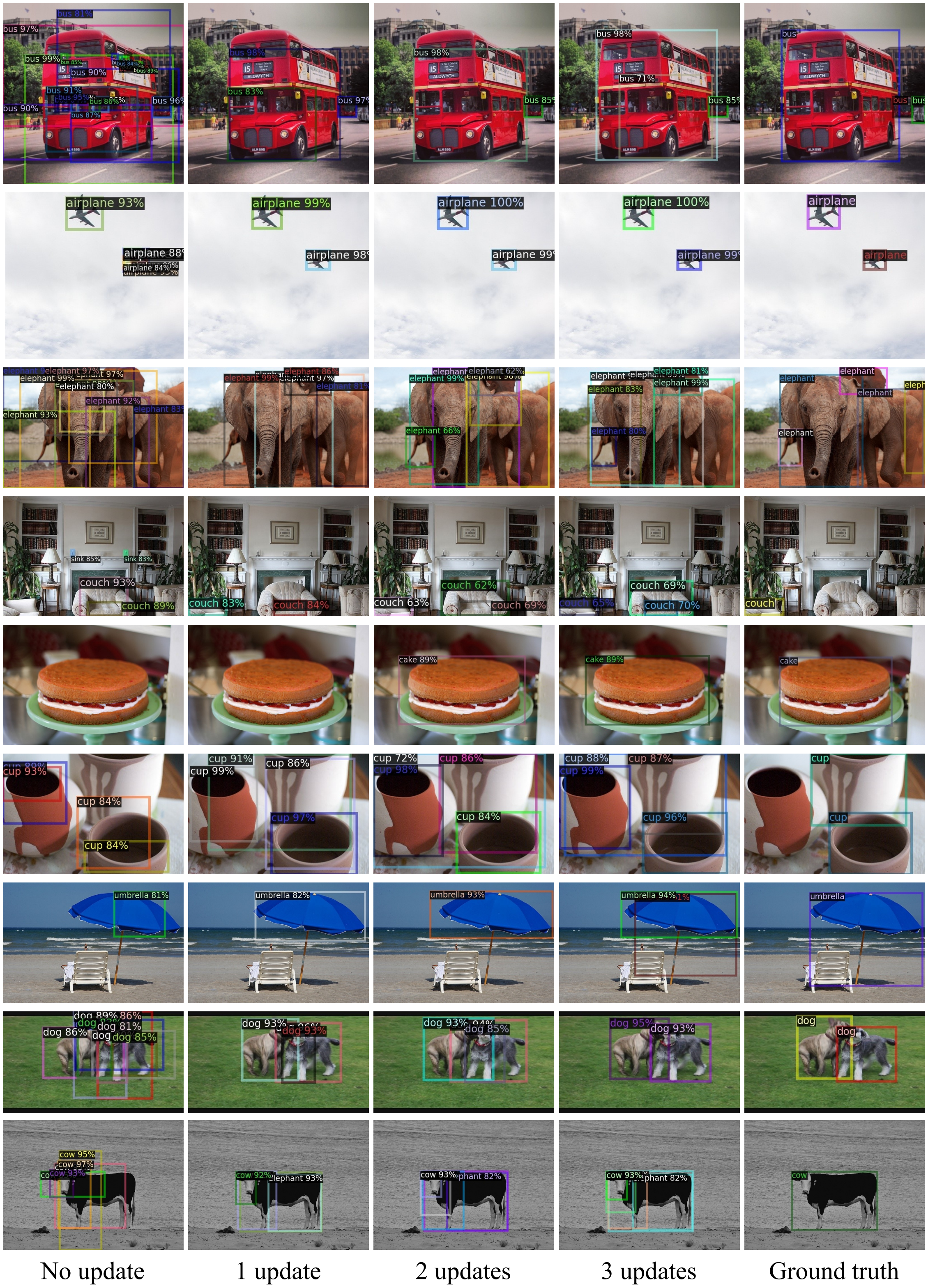


Figure 3. Visualizations of PLs with different numbers of updates for several COCO samples. Updates remove the noise and improve the quality of PLs

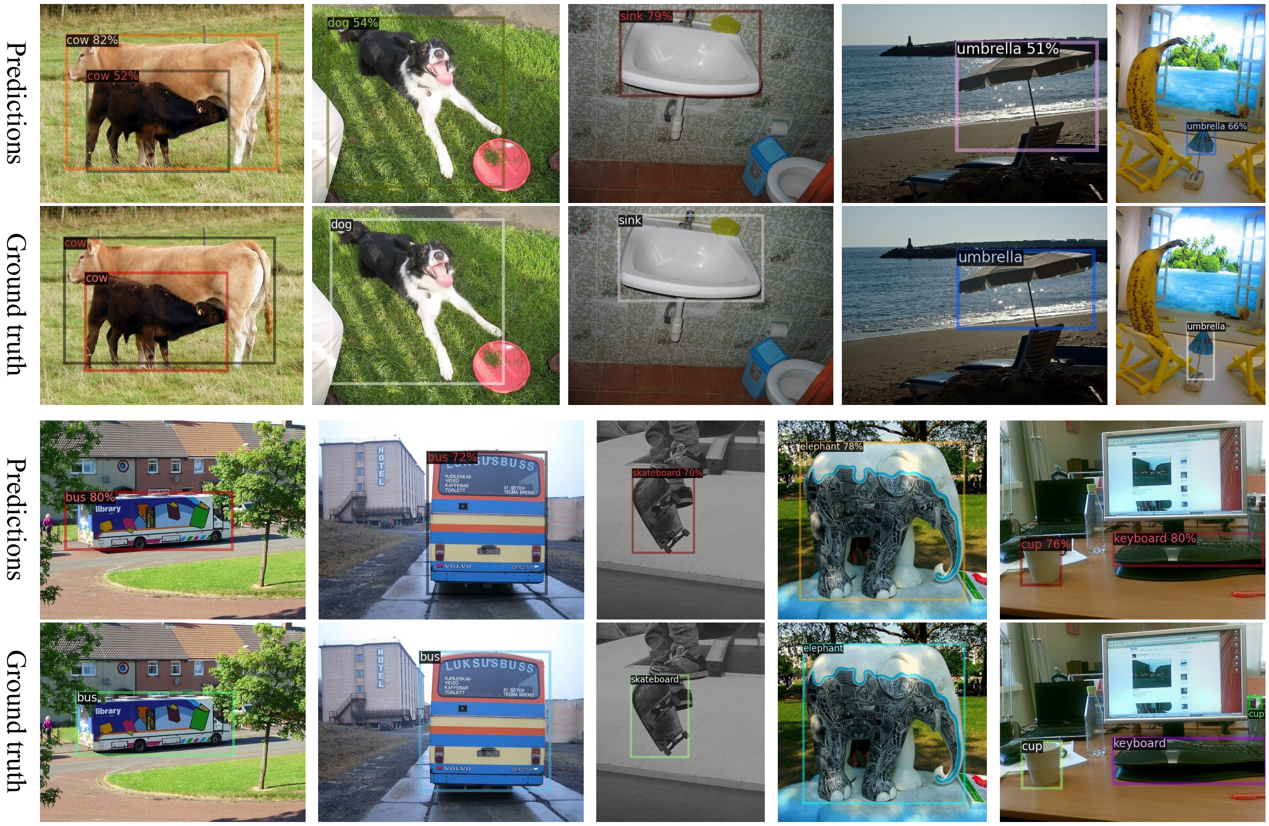
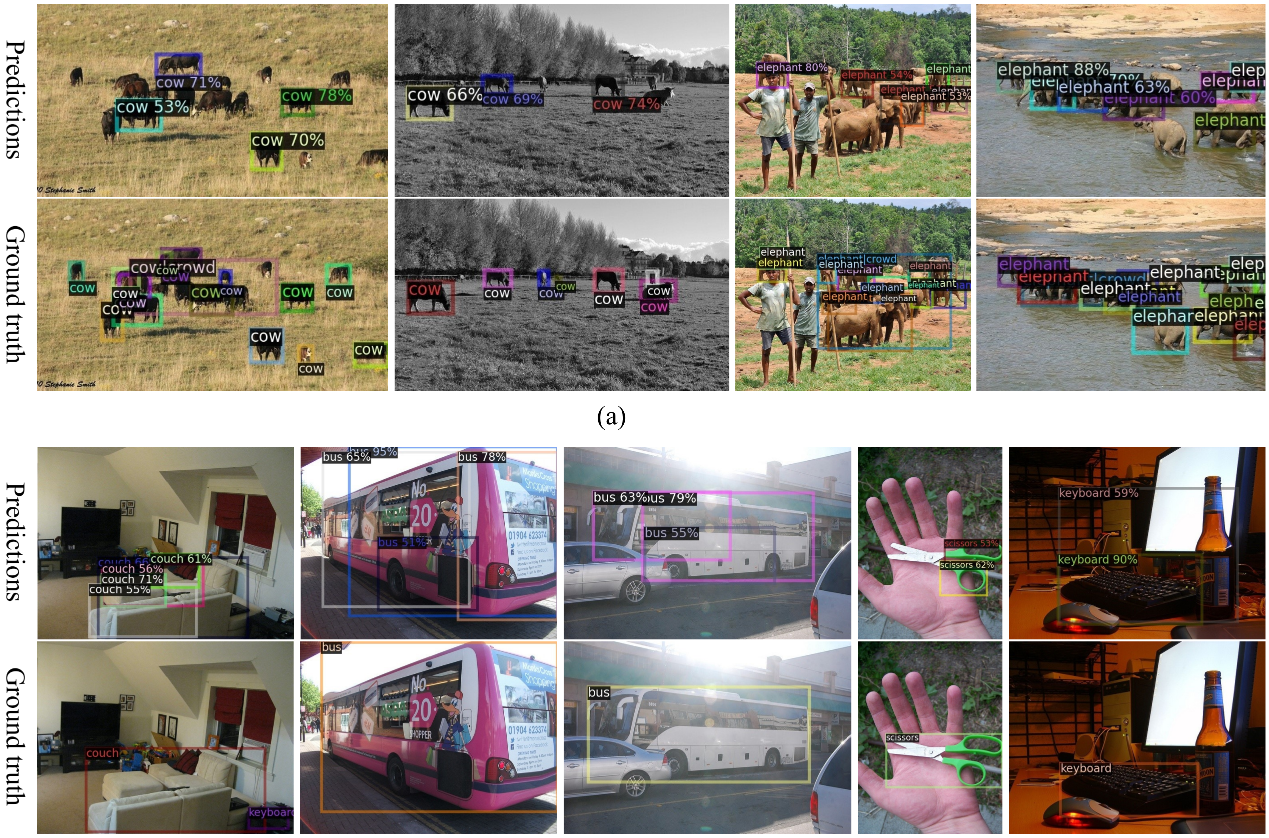


Figure 4. Good cases of the final detector on COCO. Only objects of novel categories are provided. Rare objects can be detected, e.g. a toy umbrella, a bus with rich textures, and an elephant sculpture.

[14] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, pages 159–175. Springer, 2022. 1, 3, 4

[15] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based Language-Image Pretraining. In *CVPR*, 2022. 1, 2

[16] 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



(a)

(b)

Figure 5. Failure cases of the final detector on COCO. Only objects of novel categories are provided. Two major types: **(a)** Missing instances. **(b)** Redundant predictions.