MarvelOVD: Marrying Object Recognition and Vision-Language Models for Robust Open-Vocabulary Object Detection

Kuo Wang[1], Lechao Cheng[2], Weikai Chen[3], Pingping Zhang[4], Liang Lin[1,5], Fan Zhou[1], and Guanbin Li[1,5]

[1] Sun Yat-sen University
[2] Hefei University of Technology
[3] Tencent America
[4] Dalian University of Technology
[5] Peng Cheng Laboratory

In this supplemental materials, we present more ablation studies in Section A, more comparison and compatibility analysis with existing OVD works in Secton B, implementation details in Section C, and limitations in Section D.

A Further Ablation Studies

A.1 Different indicators for adaptive reweighting

We examine several other measurements to estimate the reliability score r_i for adaptive reweighting, including:

- *Confidence of pseudo label ($r_i = s_i$)*: For comparsion. By setting $r_i = s_i$, the adaptive reweighting degenerates to conventional training design [6] with weighted pseudo-labels [16].
- *Intersection-over-Union ($r_i = iou_i$)*: Boxes with bigger overlaps from pseudo-label get larger weights and vice versa.
- *Novelty estimation ($r_i = s_i^{det}$)*: s_i^{det} defined in Eq.3 in main paper estimates the probability of containing real novel objects, we repeatedly use it for both pseudo-label mining and adaptive reweighting in this setting.

As demonstrated in Table 1, background score predicted with weakly augmented images achieves the best performance, and it significantly outperforms the conventional training design [6] with weighted pseudo-labels [16].

A.2 Ablation on global novel loss weight γ

Conventional methods [3, 4, 16] typically treat noisy pseudo-labels as ground truth and combine it with the base annotations to train the detector. In contrast, our proposed MarvelOVD strives to reduce the noises in both pseudo-labels and training boxes before utilizing them in training, which achieves more promising results. We examine different global novel loss weights and report the performance in Table 2. The results show that setting the global novel weight

Table 1: Performance of different measurements for reliability score in adaptive proposal reweighting.

r_i	$1 - b_i$	s_i	iou_i	s_i^{det}
AP_{50}^{Novel}	39.8	37.8	37.6	38.0

Table 2: Effects of different global novel loss weight γ.

Models	VL-PLM		MarvelOVD	
	AP_{50}^{N}	AP_{50}^{B}	AP_{50}^{N}	AP_{50}^{B}
$\gamma = 1$	32.7	54.0	37.8	57.0
$\gamma = 2$	32.5	53.9	**38.9**	56.5
$\gamma = 4$	-	-	38.6	56.0

as $\gamma = 2$ further improves the performance of our MarvelOVD while slightly degrading the baseline method. The reason is that the massive noise contained in the baseline method prevents further improvements. On the contrary, our approach effectively purifies the pseudo-labels and de-bias the following training designs by online mining and adaptive reweighting, allowing better performance with larger novel loss weights. Higher novel weights are also tested but do not contribute to better results.

A.3 Qualitative Results

In Figure 1, we visualize the effects of our proposed stratified label assignment and online object mining. A notable observation is that base boxes might be incorrectly labeled as novel objects due to overlaps between the pseudo-label and base annotation, potentially impairing the detector's performance on base categories. Our stratified matching strategy rectifies these mislabeled base boxes, enabling the model to assimilate novel concepts without diminishing its base detection capabilities. Additionally, the bottom two rows of Figure 1 demonstrate the efficacy of online object mining, where our approach, aided by the detector, effectively filters noise from CLIP-generated pseudo-labels.

A.4 Effects of the weak-strong augmentation

We apply the weak-strong augmentations that are not adopted by other OVD works. The motivation comes from recent semi-supervised learning methods (e.g. fixmatch [10], unbiased-teacher [9]), which demonstrate that enforcing the same supervision between weak-strong augmented features leads to better performance for learning on pseudo-labels. We apply weak-strong augmentations on the conventional OVD method [16] and MarvelOVD to evaluate its effect and the results are shown in Table 3. The weak-strong augmentation barely influences the average precision on base categories while equally improving the performance on

Fig. 1: Visualization of stratified label assignment and online object mining.

Table 3: Effects of Weak-Strong Augmentation (referred as WSA).

Models	VL-PLM		MarvelOVD	
	AP_{50}^{N}	AP_{50}^{B}	AP_{50}^{N}	AP_{50}^{B}
w/o WSA	32.7	54.0	37.2	56.4
w/ WSA	34.2	53.9	38.9	56.5

novel categories in both VL-PLM and our MarvelOVD. Without the augmentation, our framework still outperforms the base method by a significant margin.

B Comparisons and compatibility with existing OVD works

B.1 Performance Comparison

We mainly compare our method with other pseudo-label-based OVD works in the main paper. Table 4 demonstrates a more complete comparison between our MarvelOVD and other existing OVD works, including both transfer learning and knowledge distillation methods. Among them, our method still performs favorably against the state-of-the-art methods. Pseudo-label plays an important role in recent OVD works, where most advanced methods learn novel concepts from pseudo-labels generated by pretrained VLMs. With respect to VLM-generated

Table 4: Comparison with state-of-the-art methods on COCO2017 dataset.

Methods	Training Source	AP_{50}^{Novel}	AP_{50}^{Base}	AP_{50}^{All}
OV-RCNN [14]	box-level labels in $\mathcal{C}_B$, transfer learning with COCO-captions	22.8	46.0	39.9
VLDet [8]		32.0	50.6	45.8
LocOv [2]		28.6	51.3	45.7
Detic [18]	box-level labels in $\mathcal{C}_B$, internet sourced classification data, image-level labels for $\mathcal{C}_B \cup \mathcal{C}_N$	27.8	47.1	45.0
ViLD [5]	box-level labels in $\mathcal{C}_B$, knowledge distillation from CLIP	27.6	59.5	51.3
BARON [12]		34.0	60.4	53.5
RegionCLIP [17]	box-level labels in $\mathcal{C}_B$, internet sourced image-text pairs, pretraining with pseudo box-level labels	31.4	57.1	50.4
Gao *et al.* [4]	box-level labels in $\mathcal{C}_B$, internet sourced image-text pairs, pseudo-box labels in $\mathcal{C}_N$ generated by ALBEF	30.8	46.1	42.1
PromptDet [3]	box-level labels in $\mathcal{C}_B$, internet sourced image-text pairs, pseudo-box labels in $\mathcal{C}_N$ generated by CLIP	26.6	-	50.6
OADP [11]	box-level labels in $\mathcal{C}_B$, knowledge distillation from CLIP, pseudo-box labels in $\mathcal{C}_N$ generated by CLIP	35.6	55.8	50.5
Rasheed *et al.* [1]	box-level labels in $\mathcal{C}_B$, pseudo-box labels in $\mathcal{C}_N$ internet sourced image-text pairs, image-level labels for $\mathcal{C}_B \cup \mathcal{C}_N$	36.6	54.0	49.4
SAS-Det [15]	box-level labels in $\mathcal{C}_B$, pseudo-box labels in $\mathcal{C}_N$ generated by roi-align from CNN-based-CLIP	<u>37.4</u>	58.0	53.0
VL-PLM [16]	box-level labels in $\mathcal{C}_B$, box-level pseudo-labels in $\mathcal{C}_N$ generated with CLIP	32.3	54.0	48.3
MarvelOVD(Ours)		**38.9**	56.5	51.9

pseudo labels, our MarvelOVD identifies the root causes of its noises and proposes the dedicated noise-removal strategy by integrating the context-sensing capability of the detector, which consistently improves the recent advanced method by significant margins.

In particular, extracting CLIP embedding by roi-align from CNN-based CLIP backbones has been exploited in recent studies [7, 13], which provides a substituting context-aware operation for pseudo-label generation. Even though, our approach still stably outperforms methods that exploit such operation (e.g. SAS-Det [15] that recently published on arxiv). The result further indicates the effectiveness of our MarvelOVD in de-noising the pseudo-label-based learning paradigms.

Table 5: Detail of data augmentations. Probability in the table indicates the probability of applying the corresponding image process.

Weak Augmentation		
Process	Probability	Parameters
Horizontal Flip	0.5	None
Strong Augmentation		
Process	Probability	Parameters
Color Jittering	0.8	(brightness, contrast, saturation, hue) = (0.4, 0.4, 0.4, 0.1)
Grayscale	0.2	None
GaussianBlur	0.5	(sigma_x, sigma_y) = (0.1, 2.0)
CutoutPattern1	0.7	scale=(0.05, 0.2), ratio=(0.3, 3.3)
CutoutPattern2	0.5	scale=(0.02, 0.2), ratio=(0.1, 6)
CutoutPattern3	0.3	scale=(0.02, 0.2), ratio=(0.05, 8)

C More Implementation Details

C.1 Weak-Strong augmentations

The detailed weak-strong augmentations adopted in our method are illustrated in Table 5, which is identical with the semi-supervised object detection work unbiased teacher [9]. We only apply it to the COCO dataset. The augmentation on the LVIS dataset follows the common CenterNet2 benchmark [18].

C.2 Candidate pseudo-label assignment

We follow the pseudo-label-generation pipeline in VL-PLM to assign candidate pseudo-labels to each image before the training. In particular, the class-agnostic proposal generator is actually a detector trained with the base annotation (regarding all the base annotations as one class). Then we infer the train image with the proposal generator and recursively refine the predicted boxes with the RoI head by 10 times. The recursive refinement improves the localization quality of the boxes. After gathering the candidate regions, the prediction probability distribution $\boldsymbol{p}_i$ for each box is encoded by CLIP ViT-B/32 as follows:

$$\begin{aligned} \boldsymbol{r}_i &= \phi(E_{img}(R_i^{1\times}) + E_{img}(R_i^{1.5\times})) \\ \boldsymbol{p}_i &= softmax\{\boldsymbol{r}_i \cdot E_{txt}(NovelCategories)^T\} \end{aligned} \tag{1}$$

E_{img}, E_{txt} is the image-encoder and text-encoder of CLIP, $R_i^{1\times}$ is the box produced by proposal generator and $R_i^{1.5\times}$ is a region cropped by $1.5\times$ the size of $R_i^{1\times}$. After getting the probability distribution for each box, we filter them with a threshold 0.5 and post-process the remaining with NMS to obtain the candidate pseudo-labels. We record the TOP-1 CLIP score and the predicted category of the candidates and assign them to the image, and then we dynamically select

reliable ones for training under the guidance of detector. The threshold 0.5 is not a special hyperparameter that influences the performance, it's used to remove the redundant boxes that would never be selected as pseudo-labels, which accelerates the training speeds. Refining the localization with the RoI head and extracting the region-embedding with a larger area are existing techniques that adopted by the base method VL-PLM. We also maintain them in our candidate pseudo-label assignment process.

D Limitations

Our proposed MarvelOVD provides a better measurement to purify pseudo-labels from the fixed candidate boxes. It can not promote the localization quality of the pseudo-label. Since the candidate boxes are produced by a proposal generator trained with only base annotations, the localization quality for novel objects is limited. As the detector gains more knowledge of the novel object through pseudo-labels during training, its ability to localize the novel object should also be enhanced. How to rationally utilize the detector to dynamically optimize the localization quality of pseudo-labels is worthwhile exploring in the future.

References

1. Bangalath, H., Maaz, M., Khattak, M.U., Khan, S.H., Shahbaz Khan, F.: Bridging the gap between object and image-level representations for open-vocabulary detection. Advances in Neural Information Processing Systems **35**, 33781–33794 (2022) 4
2. Bravo, M.A., Mittal, S., Brox, T.: Localized vision-language matching for open-vocabulary object detection. In: DAGM German Conference on Pattern Recognition. pp. 393–408. Springer (2022) 4
3. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: European Conference on Computer Vision. pp. 701–717. Springer (2022) 1, 4
4. Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Open vocabulary object detection with pseudo bounding-box labels. In: European Conference on Computer Vision. pp. 266–282. Springer (2022) 1, 4
5. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021) 4
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 1
7. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639 (2022) 4
8. Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., Cai, J.: Learning object-language alignments for open-vocabulary object detection. arXiv preprint arXiv:2211.14843 (2022) 4
9. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021) 2, 5

10. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020) 2
11. Wang, L., Liu, Y., Du, P., Ding, Z., Liao, Y., Qi, Q., Chen, B., Liu, S.: Object-aware distillation pyramid for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11186–11196 (2023) 4
12. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15254–15264 (2023) 4
13. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. Advances in Neural Information Processing Systems **36** (2024) 4
14. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021) 4
15. Zhao, S., Schulter, S., Zhao, L., Zhang, Z., G, V.K.B., Suh, Y., Chandraker, M., Metaxas, D.N.: Taming self-training for open-vocabulary object detection (2023) 4
16. Zhao, S., Zhang, Z., Schulter, S., Zhao, L., Vijay Kumar, B., Stathopoulos, A., Chandraker, M., Metaxas, D.N.: Exploiting unlabeled data with vision and language models for object detection. In: European Conference on Computer Vision. pp. 159–175. Springer (2022) 1, 2, 4
17. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022) 4
18. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: European Conference on Computer Vision. pp. 350–368. Springer (2022) 4, 5