LOVD: Large-and-Open Vocabulary Object Detection

Anonymous Authors

1 BUILDING LARGE VOCABULARY

The number of object categories in both COCO and LVIS is limited for large-and-open vocabulary evaluation. To augment the existing vocabulary for these datasets, we apply two principled criteria: (1) The newly added categories must not overlap or be synonymous with any categories of vocabulary in the validation set. (2) They should be absent from all images in the validation set.

We initiate our selection by extracting over 10,000 categories from the OpenImages and ImageNet-21K databases as our starting point. These categories are encoded together with the vocabulary from the validation set using the CLIP text encoder to filter out those with significant semantic similarities. A multi-modal large language model then examines the validation set images with prompts like 'Is there a <category>?'. Categories that are positively identified undergo a manual re-examination and are discarded if confirmed. The final step involves a thorough manual refinement to create a refined list of 2,000 categories.

2 MORE RESULTS ON LVIS

The quantitative results of RegionCLIP, enhanced by integrating LOVD, are detailed in Table 1. This integration on the LVIS benchmark dataset results in substantial performance enhancements for RegionCLIP.

Method	**LOV**	**OV**	**DR (%)**
RegionCLIP [5] (RN50)	12.9	14.2	91.5
RegionCLIP [5] (RN50) + LOVD	15.5	16.3	95.1
RegionCLIP [5] (RN50x4)	17.9	19.7	91.0
RegionCLIP [5] (RN50x4) + LOVD	21.1	22.0	96.0

Table 1: More results on the LVIS dataset for large-and-open vocabulary setting. We report mAP as the evaluation metric.

3 SCALE OF VOCABULARY

We show more detail result of introducing a variable size of large vocabulary on model performance, as depicted in Table 2. As we incrementally increase the count of extra categories from 0 to 2000, the performance of OVD methods consistently worsens. However, our LOVD model exhibits negligible impact, indicating superior adaptability and robustness in handling the increase of potentially distracting categories in real-world scenarios.

Method	Number of Categories				
	0	500	1000	1500	2000
PBOVD [2]	30.8	24.2	23.0	20.7	19.9
VLDet [3]	32.0	28.2	26.9	25.7	24.7
RegionCLIP [5]	39.1	26.7	20.4	17.3	16.1
Detic [6]	27.8	26.8	26.1	25.7	25.4
Detic [6] + SIC-CADS [1]	26.2	25.6	25.5	25.3	25.1
Detic [6] + LOVD	27.3	27.2	26.8	26.5	26.6
CORA [4] (RN50)	41.7	35.5	34.2	32.7	32.1
CORA [4] (RN50) + LOVD	38.9	37.8	37.1	36.7	36.4
CORA [4] (RN50x4)	41.7	35.5	34.2	32.7	32.1
CORA [4] (RN50x4) + LOVD	**43.3**	**41.2**	**40.8**	**40.5**	**40.3**

Table 2: Impact of number of categories on the COCO dataset. We report AP50 as the evaluation metric.

REFERENCES

[1] Ruohuan Fang, Guansong Pang, and Xiao Bai. 2024. Simple image-level classification improves open-vocabulary object detection. In *AAAI*. 1716–1725.

[2] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. 2022. Open vocabulary object detection with pseudo bounding-box labels. In *ECCV*. 266–282.

[3] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. 2022. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint* (2022).

[4] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. CORA: Adapting CLIP for Open-Vocabulary Detection with Region Prompting and Anchor Pre-Matching. In *CVPR*. 7031–7040.

[5] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *CVPR*. 16793–16803.

[6] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *ECCV*. 350–368.