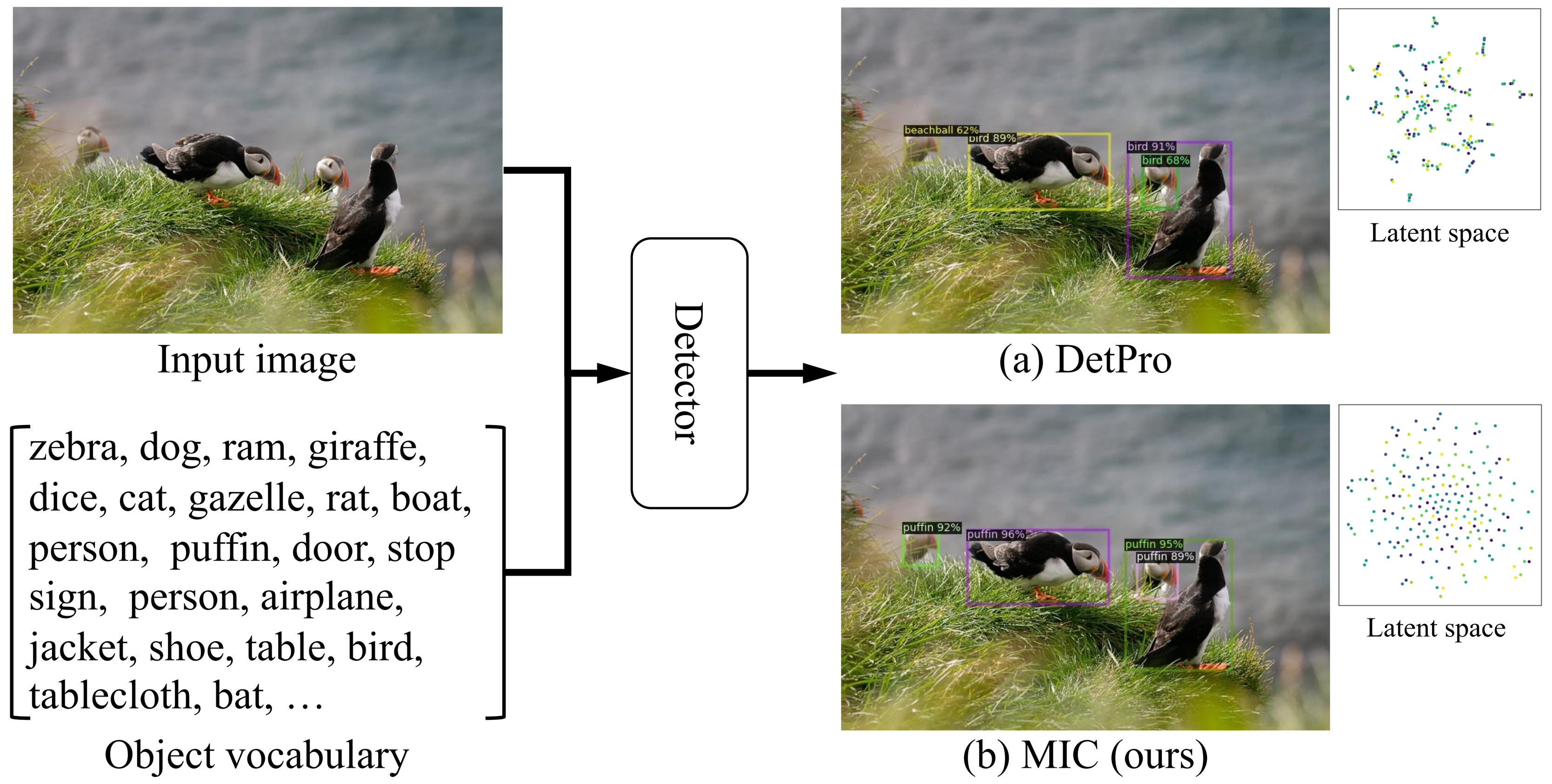
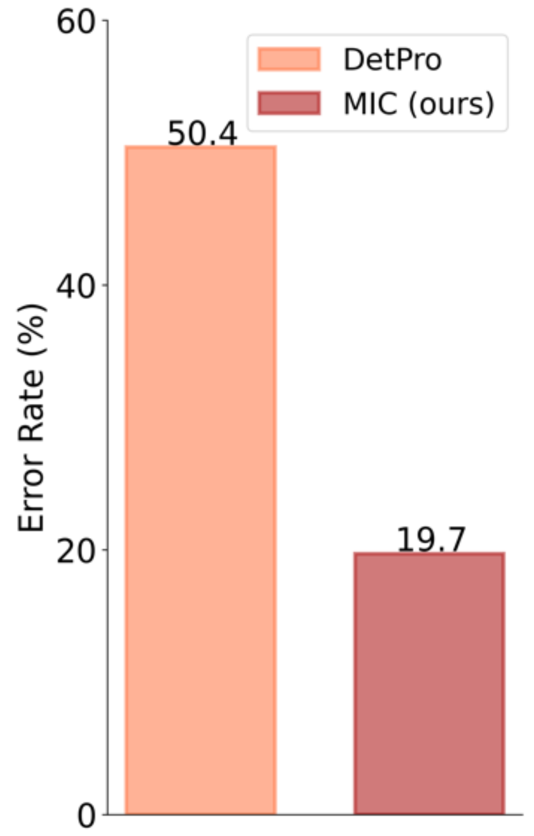
Open-Vocabulary Object Detection with Meta Prompt Representation and Instance Contrastive Optimization

Zhao Wang[1], Aoxue Li[2], Fengwei Zhou[2], Zhenguo Li[2], Qi Dou[1]
[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China [2] Huawei, China

Introduction

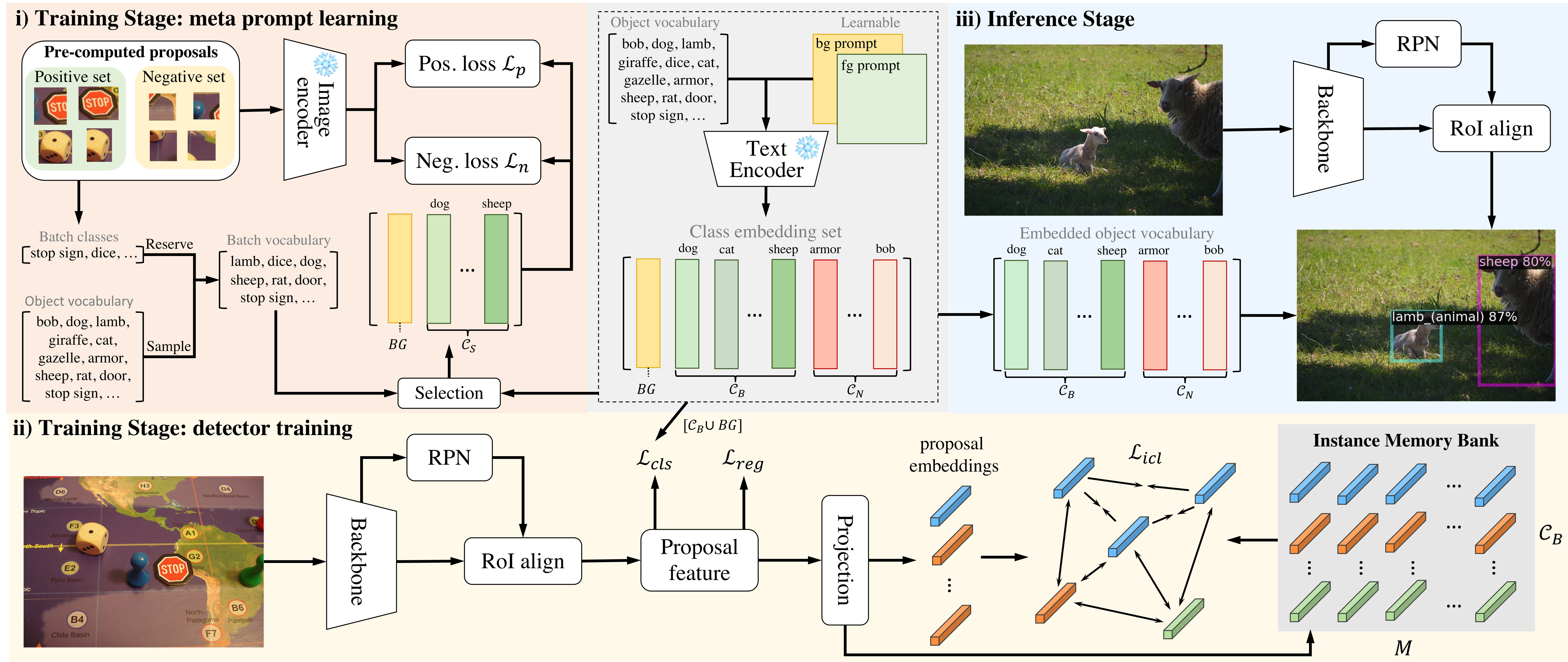
Setting & Challenge: In OVOD [1], the detector aims to detect any objects within an object vocabulary in an input image. Previous method, e.g., DetPro [2], can easily misclassify some highly similar classes (puffin v.s. bird).





Contribution: We propose a meta prompt and instance contrastive learning strategy to improve the model generalization ability, which can be more discriminative to these similar categories.

Method



- Meta prompt learning scheme to simulate a novel-class-emerging scenario
- Instance-level contrastive strategy for intra-class compactness and inter-class separation

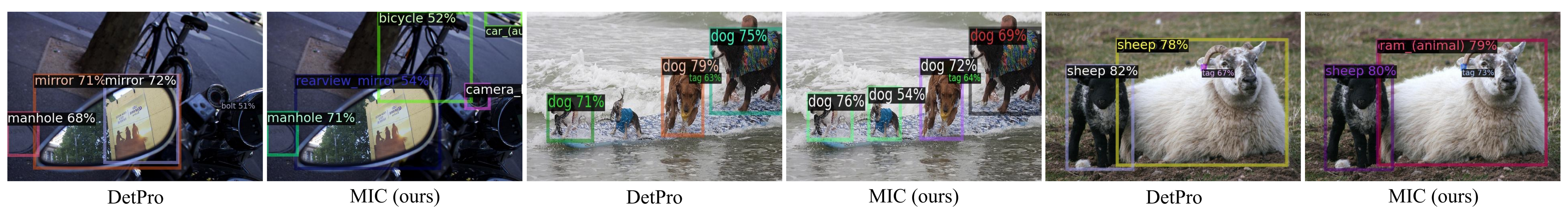
Main Results

Method	KD?	Ens?	Extra data?	Detection AP_r	AP_c	AP_f	AP	Instance segmentation AP_r	AP_c	AP_f	AP
ViLD [11]	yes	yes	no	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
RegionCLIP [38]	no	no	CC3M	17.1	27.4	34.0	28.2	-	-	-	-
DetPro [6]	yes	yes	no	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
OV-DETR [37]	yes	no	no	-	-	-	-	17.4	25.0	32.5	26.6
PromptDet [8]	no	no	LAION-400M	-	-	-	-	19.0	18.5	25.8	21.4
Detic [42]	no	no	CC3M	-	-	-	-	19.8	-	-	31.0
Rasheed *et al.* [1]	yes	no	ImageNet21k	-	-	-	-	19.3	23.6	27.9	24.1
MIC (ours)	no	no	no	**22.1**	33.9	40.0	33.8	**20.3**	30.6	35.2	30.6
MIC* (ours)	no	no	100 class names	**22.9**	34.0	39.9	34.4	**20.8**	30.5	35.4	30.7

Comparison of our method with previous SOTA methods on LVIS benchmark

Method	Pascal VOC AP_{50}	AP_{75}	COCO AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	Objects365 AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Supervised	78.5	49.0	46.5	67.6	50.9	27.1	67.6	77.7	25.6	38.6	28.0	16.0	28.1	36.7
ViLD [11]	73.9	57.9	34.1	52.3	36.5	21.6	38.9	46.1	11.5	17.8	12.3	4.2	11.1	17.8
DetPro [6]	**74.6**	57.9	34.9	53.8	37.4	22.5	39.6	46.3	12.1	18.8	12.9	4.5	11.5	18.6
MIC (ours)	73.0	**58.3**	**39.2**	**56.8**	**42.2**	**27.2**	**43.1**	**51.1**	**14.0**	**20.1**	**15.2**	**6.6**	**16.6**	**24.6**

Comparison of our method with previous SOTA methods on transfer experiments



Qualitative detection visualization results of our proposed method MIC and DetPro

Ablation

Prompt FG	Prompt BG	Strategy MPL	Strategy ICL	Detection AP_r	AP_c	AP_f	AP
fixed	✗	✗	✗	17.6	34.4	40.2	33.8
learnable	✗	✗	✗	19.7	34.0	39.8	33.8
learnable	✗	✓	✗	20.6	33.5	39.8	33.7
learnable	learnable	✓	✗	21.2	34.0	39.9	34.1
learnable	learnable	✓	✓	**22.1**	33.9	40.0	34.2

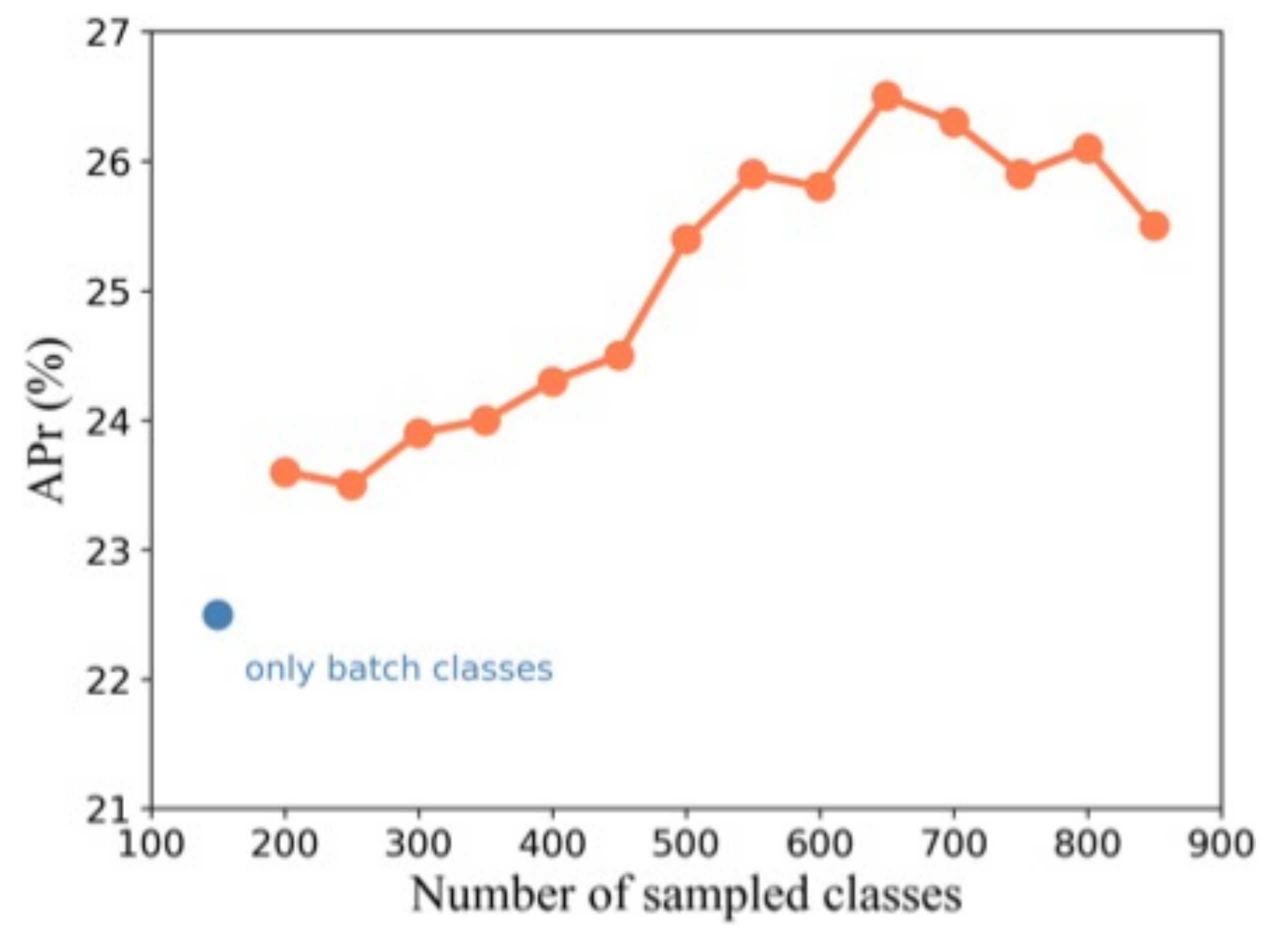
(a) Components Analysis

$[L_p, L_n]$	[4, 6]	[8, 10]	[16, 18]
AP_r	25.2	**26.4**	25.8
AP	39.3	40.1	39.7

(b) Context lengths

Position	Front	Middle	End
AP_r	23.8	25.4	**26.4**
AP	39.0	39.8	40.1

(c) Different positions of [CLS]



(d) Sampling strategy in MPL

Conclusion

We propose a novel framework MIC for open-vocabulary object detection by simulating a novel-class-emerging scenario and expanding the low-density regions in the latent feature space. Without complex training techniques and extra training data, extensive experimental results show the strong generalization ability of our proposed method.

Reference

[1] Gu, Xiuye, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. "Open-vocabulary Object Detection via Vision and Language Knowledge Distillation." ICLR, 2022.
[2] Du, Yu, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. "Learning to prompt for open-vocabulary object detection with vision-language model." CVPR, 2022.