Scene-adaptive and Region-aware Multi-modal Prompt for Open Vocabulary Object Detection

Supplementary Material

1. Training and Inference

In the training stage, the knowledge of the VLM is adapted by distilling features extracted from the visual encoder to the detector, and align with the semantic space of the classifier. Therefore, we distill features from the whole image as well as cropped region features. Specifically, for the feature $\mathbf{F}$ extracted from image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we obtain the feature representation of the whole image in the student network as follows:

$$\boldsymbol{e}_g = h(\text{GAP}(\mathbf{F})) \tag{1}$$

where GAP is the average pooling, and h is the projection layer.

Then in the teacher work, we obtain the $\bar{\mathcal{E}}_g = \{\bar{e}_g\}$ from the visual encoder $\mathcal{V}$, then the global distillation loss $\mathcal{L}_G$ is:

$$\mathcal{L}_G = \mathcal{L}_1(\mathcal{E}_g, \bar{\mathcal{E}}_g) \tag{2}$$

where $\mathcal{L}_1$ is the L_1 loss.

For the object-level distillation, we use the loss as described in paper:

$$\mathcal{L}_O = \mathcal{L}_1(\mathcal{E}_o, \tilde{\mathcal{E}}'). \tag{3}$$

where $\tilde{\mathcal{E}}' = \{\tilde{e}'_p\}$ is obtained by incorporating the region prompts and text prompts.

In the testing stage, we calibrate the prediction score as previous distillation-based OVD methods [4, 12]:

$$P^O_{\mathcal{C}}(p,c) = \frac{\exp(\frac{\bar{e}_p \cdot \boldsymbol{t}_c}{||\bar{e}_p|| \cdot ||\boldsymbol{t}_c||})}{\sum_{c' \in \mathcal{C}^B \cup \mathcal{C}^N} \exp(\frac{\bar{e}_p \cdot \boldsymbol{t}_{c'}}{||\bar{e}_p|| \cdot ||\boldsymbol{t}_{c'}||})} \tag{4}$$

where $\bar{e}_p$ is the proposal embeddings extracted from the teacher network.

The calibrated probability $P'_{\mathcal{C}}(p,c)$ is:

$$P'_{\mathcal{C}}(p,c) = \begin{cases} (P_{\mathcal{C}}(p,c))^{\lambda} \cdot (P^O_{\mathcal{C}}(p,c))^{1-\lambda}, c \in \mathcal{C}^B \\ (P_{\mathcal{C}}(p,c))^{1-\lambda} \cdot (P^O_{\mathcal{C}}(p,c))^{\lambda}, c \in \mathcal{C}^N \\ 1 - \sum_{c'} P_{\mathcal{C}}(p,c'), c = bg \end{cases} \tag{5}$$

where λ is set to 2/3. Note that the distillation modules are not used during the inference phase.

2. More Experiments

2.1. Benchmark Results

We similarly added the proposed approach to the DETR-based framework, and then fused our scene adaptive prompt generator and region-aware multi-modal alignment module into it. The results on OV-COCO and OV-LVIS can be seen in Table 1 and 2.

Compared to RCNN-based methods, DETR-based methods generally exhibit higher performance on base classes, resulting in relatively superior performance when transitioning to new classes. Our approach shows a notable improvement in novel class performance compared to OV-DETR and Prompt-OVD, both of which also employ distillation techniques. This enhancement is attributed to the integration of general knowledge into the evaluation, while maintaining better performance on the base class. SAMP underscores the importance of scene-adaptive prompts featuring learnable visual and text prompts for tasks requiring dense alignment. It's worth noting that we leverage the fundamental vision-language model (VLM) CLIP without utilizing caption data, ensuring compatibility with other VLMs or pre-training datasets lacking caption information.

2.2. Analytical experiments

The effect of different number of selected tokens. Figure 1 illustrates the impact of varying prompt numbers on the COCO dataset, maintaining a fixed prompt length of 8. The analysis reveals that using a prompt number of 3 yields superior performance.

The effect of different components in generator. Table 3 shows the results obtained through the implementation of shared learnable prompts, scene prompts, and constricted scene prompts. The results indicate that utilizing shared prompts leads to a marginal 0.6% enhancement in novel class results compared to using learnable prompts alone. However, incorporating scene prompts results in a notable increase of 2.0% in novel class improvements. Thus, it can be concluded that the constructed scene prompts are more effective in improving knowledge transfer to the pre-trained models.

The effect of different components in region prompt. Table 4 presents an evaluation of various region prompt constructions, including a learnable prompt and a mask. The results show that the use of a mask leads to an improved performance on novel classes. This improvement is probably attributed to the mask offering more class information, thereby aiding in better adaptation to the detector. Conversely, solely using a learnable prompt proves less effective due to its absence of specific spatial information.

Method	Supervision	Backbone	Detector	Prompt	mAP_{50}^{N}	mAP_{50}^{B}	mAP_{50}
CORA [13]	CLIP	R50	D-DETR	T (cat)	35.1	35.5	35.4
SGDN [10]	Caption	R50	D-DETR	×	**37.5**	61.0	54.9
DK-DETR [7]	CLIP	R50	D-DETR	T(cat)	32.3	61.1	54.3
OV-DETR [17]	CLIP	R50-C4	D-DETR	T(cat)	29.4	61.0	53.7
Prompt-OVD [11]	CLIP	ViT-B/16	D-DETR	T (cat)	30.6	63.5	54.9
Ours	CLIP	ViT-B/16	D-DETR	SAP	36.8	**63.9**	**55.6**

Table 1. The comparision results with other methods on the OV-COCO dataset.

Method	Backbone	Detector	Teacher	Prompts	AP_r	AP_c	AP_f	AP
MEDET [3]	R50-FPN	CN2	-	T(cat)	22.4	-	-	33.4
VLDET [8]	R50-FPN	CN2	-	T(cat)	21.7	29.8	34.3	30.1
RO-ViT [5]	ViT-B/16	MaskRCNN	-	T(cat)	28.0	-	-	30.2
SGDN [10]	ResNet50	DeformableDETR	-	-	23.6	29.0	34.3	31.1
DK-DETR [7]	R50	D-DETR	CLIP	T(cat)	22.2	**32.0**	**40.2**	33.5
OWL-ViT [9]	ViT-H/18	DETR	-	T(cat)	23.3	-	-	35.3
Ours	VIT-B/16	DETR	CLIP	SAP	**28.5**	30.3	37.4	**35.6**

Table 2. The comparision results on LVIS dataset.

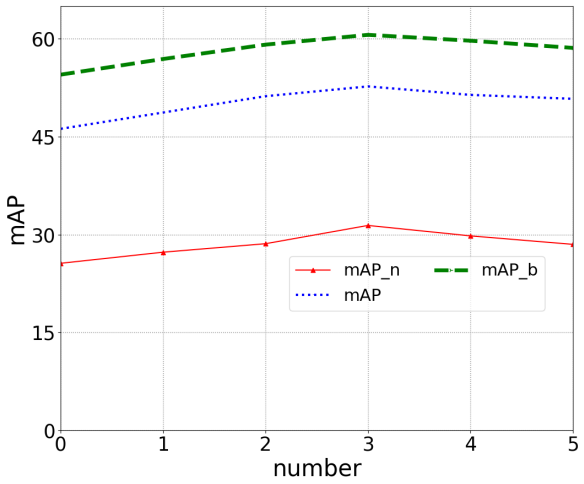

Figure 1. The effectiveness of different numbers of the selected prompts.

$\mathbf{P}^a$	$\mathbf{P}^s$	mAP_{50}^{B}	mAP_{50}^{N}	mAP_{50}
		56.5	27.6	48.2
√		57.6	28.2	49.3
	√	58.8	29.6	50.8
√	√	**59.6**	**32.8**	**53.1**

Table 3. The effectiveness of different components in generator.

learnable	mask	mAP_{50}^{B}	mAP_{50}^{N}	mAP_{50}
		56.5	27.6	48.2
√		57.2	28.3	49.6
	√	58.4	30.6	51.8
√	√	**59.3**	**32.3**	**52.7**

Table 4. The effectiveness of different prompts in region prompt.

2.3. Visulation

Detection visualization. We provide more detection results in the Figure 2. And qualitative detection results of transfer performance are shown in Figure 3. Note that the detection model is trained on the LVIS dataset and transferred to VOC and COCO datasets. It can be seen that in complex images containing more objects, the migration model can still distinguish different objects well.

The effectiveness of the region prompt. Figure 4 presents a comparison of classification accuracy using ground truth boxes between our proposed RMA and CLIP, demonstrating our improved image region inference capabilities.

We visualize activation maps from baseline and our detector in Figure 5. Taking the 2nd column as an example, the activation map of our detector accurately highlights more areas of novel objects, i.e. "cup", with our region prompt. Therefore, the region prompt in SAMP can localize objects more accurately, which further helps detect novel objects.

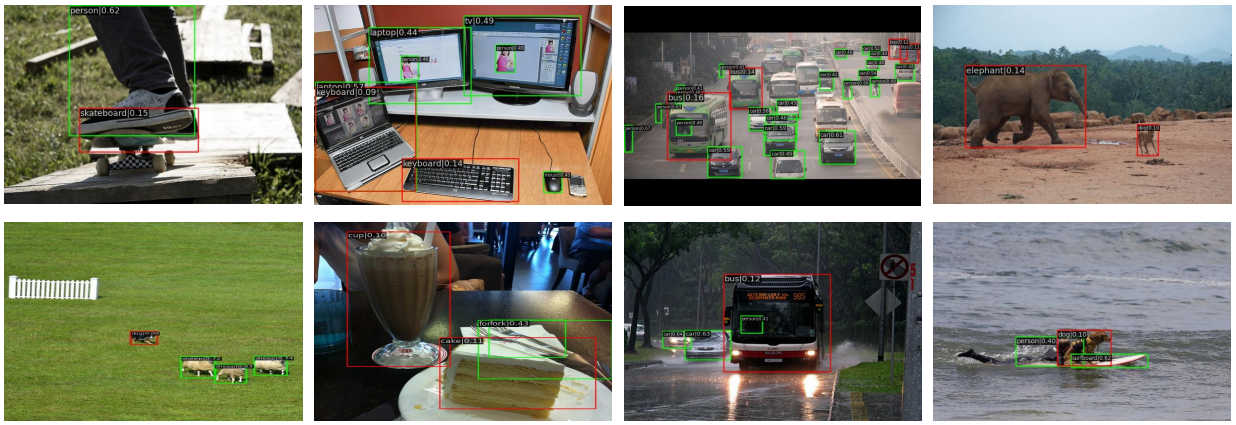

Figure 2. Visualization of the detection results.

Figure 3. Visualization of the CDE detection results on VOC and COCO dataset.

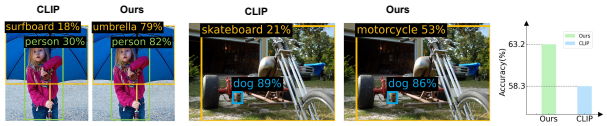

Figure 4. Classification with gt boxes on the OV-COCO.

3. Different OVD methods.

Current open vocabulary object detection (OVD) methods can be divided into four types according to the supervisory information. In this section, we will introduce the ideas and typical models of each type, and then discuss the advantages and disadvantages of the different methods.

Region-aware Training. This research focuses on efficiently aligning regions and words within inexpensive and ample image-caption pairs. By incorporating additional region-oriented losses, the models can acquire the ability to align across different modalities and expand their vocabulary. One way is to use weakly supervised grounding or contrastive loss, which uses image-text pairs to establish a coarse and noisy correspondence between regions and words. This category includes OV-RCNN [18], LocOv [2], RO-ViT [5], Detclip[15], Detclip2[16]. Another way is to leverage region-word pairs from visual grounding datasets to broaden the vocabulary. This category includes methods such as SGDN[10], MEDet[3], VLDet[8] and CORA[13].

Pseudo-Labeling. In addition to utilizing abundant image-text pairs, models that endorse pseudo-labeling incorporate large pre-trained VLMs or use self-training to generate pseudo labels. These detectors are trained on a combination of existing base annotations and newly generated pseudo labels. Depending on the type and level of detail of the pseudo labels, these methods can be categorized into pseudo region-caption pairs, region-word pairs, and pseudo captions. This approach can be considered a more effective way of prompting compared to the template prompts used in CLIP. The typical methods are Detic[20] and 3Ways[1], RegionCLIP[19], and PromptDet[11].

Knowledge Distillation-Based. Contrastively trained VMLs demonstrate enhanced zero-shot recognition capabilities in a range of subsequent tasks. Within this cat-

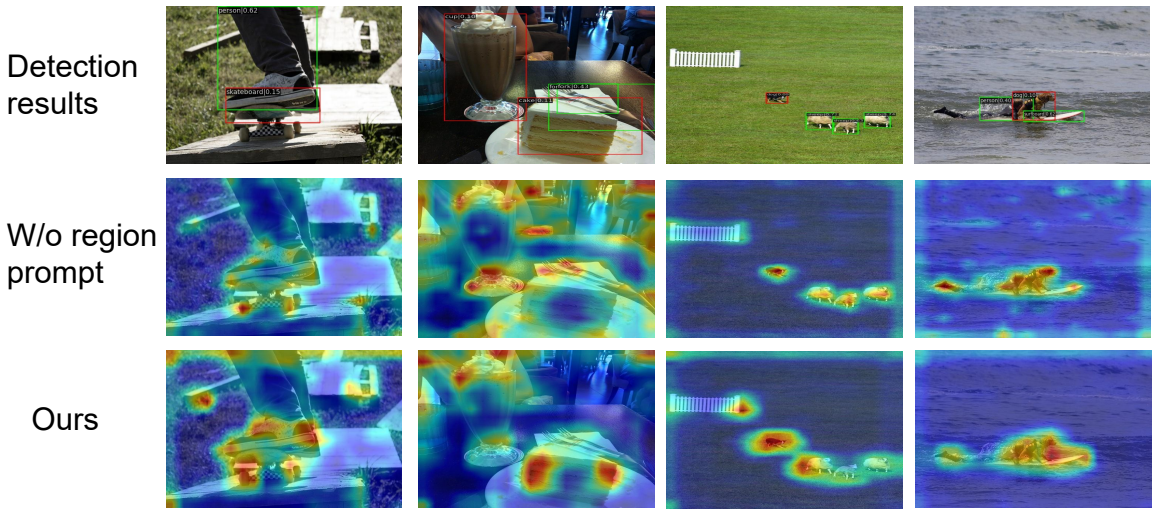

Figure 5. Visualization of the CDE detection results on VOC dataset.

egory, techniques primarily condense region embeddings from the teacher model (VMLs image encoder) into the student model, enabling compatibility with text embeddings of VMLs using detection data. It can be categorized into two subgroups: distilling region embeddings individually or collectively. In the former, individual RoIs are used as input, while in the latter, a bag of RoIs is fed into the VLMs image encoder. This type of methods consist of ViLD [4], DetPro[4], OADP[12], ZSD-YOLO[14], OV-DETR[17], Prompt-OVD[11], and BARON[2].

Transfer Learning-Based. Transfer learning-based models differ from KD-based methodologies in their utilization of VLMs. Specifically, they primarily employ the VLMs image encoder as a feature extractor. For instance, this can involve directly fine-tuning it on detection data or extracting visual features using the frozen image encoder of VLMs. This category encompasses methods OWL-ViT[9] and F-VLM[6].

Although the region-aware training and pseudo-labelling methods use relatively cheap and large number of image-text pairs, how to reduce the negative effect of noisy pairs is crucial for improving data efficiency. For knowledge distillation and transfer learning based models, the context mismatch prevents the full potential of VLMs. Pre-training images from CLIP are full images, while proposals contain limited spatial cues, the gap between image-level pre-training and region-level detection should be eliminated.

References

[1] Relja Arandjelović, Alex Andonian, Arthur Mensch, Olivier J Hénaff, Jean-Baptiste Alayrac, and Andrew Zisserman. Three ways to improve feature alignment for open vocabulary detection. *arXiv preprint arXiv:2303.13518*, 2023. 3

[2] Maria A Bravo, Sudhanshu Mittal, and Thomas Brox. Localized vision-language matching for open-vocabulary object detection. In *DAGM German Conference on Pattern Recognition*, pages 393–408. Springer, 2022. 3, 4

[3] Peixian Chen, Kekai Sheng, Mengdan Zhang, Mingbao Lin, Yunhang Shen, Shaohui Lin, Bo Ren, and Ke Li. Open vocabulary object detection with proposal mining and prediction equalization. *arXiv preprint arXiv:2206.11134*, 2022. 2, 3

[4] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1, 4

[5] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11144–11154, 2023. 2, 3

[6] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *The Eleventh International Conference on Learning Representations, ICLR*, 2023. 4

[7] Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6501–6510, 2023. 2

[8] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *The Eleventh International Conference on Learning Representations, ICLR*, 2023. 2, 3

[9] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*. 2, 4

[10] Hengcan Shi, Munawar Hayat, and Jianfei Cai. Open-vocabulary object detection via scene graph discovery. *arXiv preprint arXiv:2307.03339*, 2023. 2, 3

[11] Hwanjun Song and Jihwan Bang. Prompt-guided transformers for end-to-end open-vocabulary object detection. *arXiv preprint arXiv:2303.14386*, 2023. 2, 3, 4

[12] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2023. 1, 4

[13] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7031–7040, 2023. 2, 3

[14] Johnathan Xie and Shuai Zheng. Zero-shot object detection through vision-language embedding alignment. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1–15. IEEE, 2022. 4

[15] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 3

[16] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023. 3

[17] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 2, 4

[18] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 3

[19] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 3

[20] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 3