Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model

Method	AP_r	AP_c	AP_f	AP
ViLD-text* [1]	12.1	24.2	28.9	23.9
DetPro-text	**14.2**	23.9	28.9	24.2

Table 1. We compare our DetPro-text with the ViLD-text. * denotes our re-implementation version, see Section 5.2 for the details.

Positive samples (%)	AP_r	AP_c	AP_f	AP
10	18.2	25.4	28.2	25.3
30	18.4	25.1	28.2	25.1
50	18.8	25.4	28.2	25.4
100	**19.1**	**25.4**	**28.2**	**25.4**

Table 2. Ablation study of using different number of positive samples for DetPro training.

A. More Experiments and Analysis

A Variant of DetPro. Following ViLD [1], we present a variant of DetPro named DetPro-text, and compare it with ViLD-text. In the DetPro-text, we remove the image head and only use a text head for training and inference. We use the LVIS setting as described in Section 5.3. Table 1 shows the comparison.

Using Different Number of Positive Samples for Training. We also study the effects of using different number of positive samples in DetPro training as shown in Table 2. The LVIS setting is adopted in this study. We observe that using all positive samples results in the best generalization performance on novel classes.

Accuracy of Proposal Classification. In our DetPro, we first optimize the prompt representations then feed them into CLIP text encoder to generate class embedding as classifiers of the detector. Here we report the image proposal classification accuracy on the LVIS dataset to demonstrate the effectiveness of our approach. Concretely, given a set of proposals generated by RPN, we resize each proposal to the size of 224×224 and feed it into the CLIP image encoder to extract its image embedding, then we compute the similarities between the image embedding and all class embedding to predict its class. We compare our approach with the prompt engineering. Table 3 and Table 4 show top-1 and top-5 accuracy, respectively. Remarkable improvements are observed on both base classes and novel classes, indicating that the prompt representations learned by our DetPro are also beneficial to the open-vocabulary image classification task.

Method	Base class	Novel class
Prompt engineering	20.1	17.7
DetPro	**24.4**	**21.7**

Table 3. Top-1 accuracy of proposal classification.

Method	Base class	Novel class
Prompt engineering	39.3	37.2
DetPro	**49.0**	**40.3**

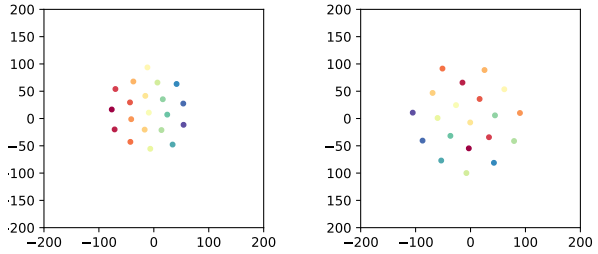
Table 4. Top-5 accuracy of proposal classification.

Assembling the Well-trained ViLD with DetPro. In Section 4.3 of the main paper, we use class embedding generated by DetPro for the ViLD training and inference. In this study, we first use class embeddings generated by prompt engineering as classifiers of the detector to train ViLD, and assemble the well-trained ViLD with our DetPro for inference, by simply replacing the original class embedding in the image head with the ones generated by our DetPro. It can be seen in Table 5 that simply assembling the original ViLD with DetPro trained with different ensemble strategies (see Table 6 of the main paper) already shows non-negligible improvements on novel classes.

T-SNE Visualization for Transferred Datasets. In Section 5.5, we generate the class embedding for the LVIS dataset and show the t-SNE figure. Here we use t-SNE to visualize the class embedding generated by our DerPro and prompt engineering on transferred datasets including Pascal VOC, COCO, and Objects365. Figure 1-3 show the comparison. We observe the same phenomenon that the class embedding generated by DetPro is more discriminative in the embedding space, which further validates their suitability serving as the region classifiers for open-vocabulary ob-

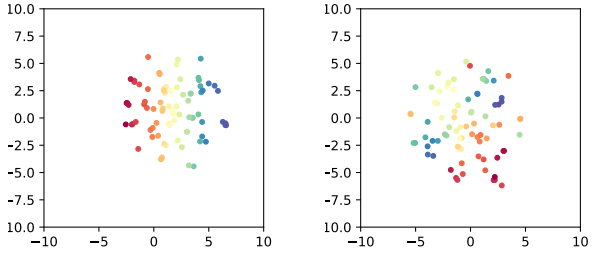
Method	AP_r	AP_c	AP_f	AP
ViLD*	16.8	25.6	28.5	25.2
DetPro-Ensemble(0.5:1.0:0.1)	18.1	25.7	28.3	25.4
DetPro-Ensemble(0.6:1.0:0.1)	18.0	25.4	28.2	25.2
DetPro-Ensemble(0.7:1.0:0.1)	18.0	25.4	28.2	25.3
DetPro-Ensemble(0.8:1.0:0.1)	17.9	25.7	28.3	25.4

Table 5. Assembling the well-trained ViLD with DetPro trained under different settings outperforms the original ViLD.



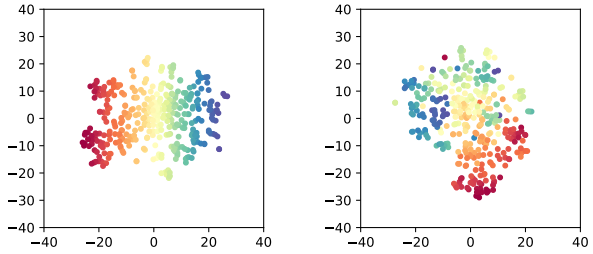
(a) Prompt engineering. (b) DetPro.

Figure 1. T-SNE visualization for Pascal VOC dataset.



(a) Prompt engineering. (b) DetPro.

Figure 2. T-SNE visualization for COCO dataset.



(a) Prompt engineering. (b) DetPro.

Figure 3. T-SNE visualization for Objects365 dataset.

ject detection.

B. More Implementation Details

More Details of Our Open-world Object Detector. We use multi-scale training with the size of (1333, 640), (1333, 672), (1333, 704), (1333, 736), (1333, 768), (1333, 800). For RPN, we apply an NMS with a threshold of 0.7 and generate a maximum of 1000 proposals. We apply a class-agnostic NMS with a threshold of 0.5 on the final predictions and set the maximum number of output bounding boxes to 300.

More Details of DetPro Training. We set the batch size as 512. We use a cross-entropy loss with a temperature parameter of 0.01.

References

[1] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.