

– Supplementary material –  
**PromptDet: Towards Open-vocabulary  
Detection using Uncurated Images**

Chengjian Feng<sup>1</sup>, Yujie Zhong<sup>1</sup>, Zequn Jie<sup>1</sup>, Xiangxiang Chu<sup>1</sup>  
Haibing Ren<sup>1</sup>, Xiaolin Wei<sup>1</sup>, Weidi Xie<sup>2, ✉</sup>, and Lin Ma<sup>1</sup>

<sup>1</sup>Meituan Inc. <sup>2</sup>Shanghai Jiao Tong University

## 1 Open-vocabulary COCO benchmark details

Standard MS-COCO [3] contains 80 categories. Following [1, 4], we only select 48 classes as base classes, and 17 classes as novel classes. The *train* set and *minival* set are the same as the standard MS-COCO 2017, however, only images containing at least one base class are used for the training. For self-training, we search for the images from LAION-400M and download about 1500 images for each novel category. We summarise the dataset statistics for open-vocabulary COCO benchmark in Table 1.

Since there are no class descriptions in the meta data of MS-COCO, we use the descriptions from LVIS if the category is covered by LVIS. For the other categories, we get their descriptions by searching on Google, for instance, {category: “donut”, description: “a small fried cake of sweetened dough, typically in the shape of a ball or ring”}.

We conduct the experiment using Mask-RCNN [2] with a ResNet-50-FPN backbone. The model is trained with a batchsize of 64 on 8 GPUs, for 24 epochs (2x learning schedule). Similar to open-vocabulary LVIS benchmark, we use COCO-base to train an initial open-vocabulary object detector. Then we use a combination of COCO-base and LAION-novel datasets for self-training. Considering the small number of the base categories on open-vocabulary COCO benchmark, we select 1000 images for each base category during the regional prompt learning, and set  $K = 10$  to retain the precise bounding boxes for the pseudo labeling. Besides, the COCO-base images contain abundant objects of the novel categories. To enhance the recall of the region proposal network (RPN) for the novel categories, we compute the similarity between the features from the negative box proposals and the text embedding of novel categories, and treat the box proposals with high similarity (*i.e.* larger than 0.1) as the positive samples, during the training of RPN.

## 2 Cross-dataset transfer

To demonstrate the generalisation, we directly evaluate the detector from open-vocabulary LVIS benchmark on MS-COCO [3], *without* any finetuning on it.

Table 1: A summary of dataset statistics for open-vocabulary COCO benchmark. The numbers in bracket refer to the number of base and novel categories.

Dataset	Train Eval.		Definition	#Images	#Categories
MS-COCO	-	-	original MS-COCO dataset	0.1M	80
LAION-400M	-	-	image-text pairs filtered by CLIP	400M	unlabeled
COCO-base	✓	✗	base categories on MS-COCO	0.1M	48
LAION-novel	✓	✗	image subset of novel categories	25K	17 (noisy)
COCO <i>minimal</i>	✗	✓	MS-COCO validation set	5K	65 (48+17)

Specifically, MS-COCO contains 80 categories, with only 59 categories covered by LVIS-base (denoted as base categories,  $\mathcal{C}_{base}$ ). Therefore, the other 21 categories can be used to benchmark the generalisation ability of PromptDet.

As shown in Table 2, the learned regional prompt improves the generalisation of the detector, and achieves 29.0 AP for the novel categories, suppressing the manual prompt by 1.6 AP (27.4 AP *vs.* 29.0 AP). Similarly, with more training (72 epochs), the model reaches 31.4 AP on detecting the 21 novel categories on MS-COCO.

Table 2: Generalisation from LVIS-base to MS-COCO with different prompts. The trained detector is directly transferred to MS-COCO by replacing the category and description embedding in the classifiers without fine-tuning.

	Prompt	Epochs	AP <sub>novel</sub>	AP <sub>base</sub>	AP
“a photo of [category], which is [description]”	manual	12	27.4	26.9	27.0
PromptDet	learnable	12	29.0	26.5	27.2
PromptDet	learnable	72	31.4	29.1	29.7

### 3 Qualitative Results

In Figure 1, we show the qualitative results from PromptDet on open-vocabulary LVIS benchmark. Without any manual annotations, the detector can now accurately localise and recognise the objects from a diverse set of novel categories.



## References

1. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: Proceedings of the European Conference on Computer Vision. pp. 384–400 (2018)
2. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the International Conference on Computer Vision. pp. 2961–2969 (2017)
3. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
4. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. arXiv preprint arXiv:2201.02605 (2022)