Learning Background Prompts to Discover Implicit Knowledge for Open Vocabulary Object Detection (Supplementary Document)

We present additional implemental details, comparison, and analysis results of the proposed LBP framework in this supplementary material.

1. Implementations.

We follow the implementation of BARON [11] to conduct experiments. Similar to [6, 11], we employ Faster R-CNN [7] coupled with ResNet50-FPN [5] as the base detector, initializing its backbone network with weights from SOCO [10]. We employ a 2× training schedule (180,000 iterations), with batch size set to 32 (16 for detection and 16 for distillation). We choose SGD [8] as the optimizer, configured with a momentum of 0.90 and a weight decay of 2.50×10^{-5}. Additionally, following BARON [11], we utilize ViT-B/32 CLIP as the PVLM model, with the fixed context prompts from ViLD [3]. For other hyper-parameters, we maintain consistency across all experiments, such as $n_a = 10$, $\theta = 0.95$ (consistent with VL-PLM [12]), $\tau = 0.02$, $\gamma = 0.02$, and $\lambda_{bg} = 0.05$.

2. Transfer to Other Datasets

Similar to [1, 11], we also report the inference performance of the proposed LBP approach and other previous state-of-the-art (SOTA) OVD methods when transferring a detector trained on the LVIS dataset to three other datasets: Pascal VOC 2007 test set [2], COCOvalidation set [4], and Objects365 v2 validation set [9]. As shown in Table 6, our LBP method demonstrates superior inference performance across these three datasets compared to existing state-of-the-art methods, showcasing the generalized applicability of our LBP approach across various scenarios.

3. Additional Ablation Analysis

Further analysis of BOD. To further illustrate the advantages of the proposed **BOD** module, we attempt to discover representative results to visualize conceptual overlaps between background underlying categories estimated during training and the novel categories during inference, demonstrated in Figure 4. To be more specific, we choose the representative proposals with high predicted probability scores for three background underlying categories.

Figure 4(a) shows one of the estimated background categories, which exhibits significant overlap with “bus” proposals belonging to the novel categories. This observation strongly demonstrates the effect of our proposed **BOD** in discovering meaningful latent categories from a multitude of background proposals during model training.

Conversely, the other estimated category portrayed in Figure 4(b) demonstrates substantial connections with objects associated with two distinct novel categories: “skateboard” and “snowboard”. This suggests that while **BOD** might not precisely differentiate all novel categories, it effectively leverages knowledge from visually similar-looking categories, thereby significantly enhancing the model’s representations of objects w.r.t. those categories.

In essence, both Figure 4(a) and Figure 4(b) distinctly illustrate conceptual overlaps in contextual embeddings between the background underlying categories estimated during training and the novel categories accessed during inference.

Moreover, Figure 4(c) illustrates that the represented background underlying category encompasses diverse types of food, devoid of significant semantic overlap with the inference novel categories, and nevertheless, it can still detect objects pertinent to that category. This discovery underscores the substantial capacities of the proposal model in discovering and leveraging the knowledge of implicit objects from background proposals, markedly bolstering feature discrimination output by our model, and consequently, significantly enhancing detector performance.

More analysis of IPR. To further validate the necessity of **IPR**, we visualize the distributions of the contextual embeddings, encoded by text encoder of CLIP, from base categories $\mathcal{C}_b$, novel categories $\mathcal{C}_u$, and background underlying categories $\mathcal{C}_O$ in the OV-COCO task, depicted in Figure 5. As illustrated, during inference, several embeddings from $\mathcal{C}_O$ closely resemble those of some novel categories from $\mathcal{C}_u$, indicating a probable semantic similarity or conceptual overlap between the two categories. As displayed in Eq. (16), for each $c' \in \mathcal{C}_O$, the proposed **IPR** module adjusts

	Pascal VOC		MS-COCO						Objects365					
	AP$_{50}$	AP$_{75}$	AP	AP$_{50}$	AP$_{75}$	AP$_s$	AP$_m$	AP$_l$	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§ [3]	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§ [1]	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
BARON‡ [11]	76.0	58.2	36.2	55.7	39.1	24.8	40.2	47.3	13.6	21.0	14.5	5.0	13.1	20.7
LBP‡ (ours)	**76.1**	**58.4**	**36.8**	**56.5**	**39.8**	**25.6**	**40.6**	**48.1**	**14.3**	**21.8**	**15.1**	**5.5**	**13.7**	**21.6**

Table 6. Comparison results of LBP and existing SOTA methods on Pascal VOC test set, COCO validation set and Object365 validation set with the model being trained on OV-LVIS. Specifically, § indicates those results reported from DetPro [1], while ‡ indicates the model trained using learnable prompt templates, proposed by DetPro [1].

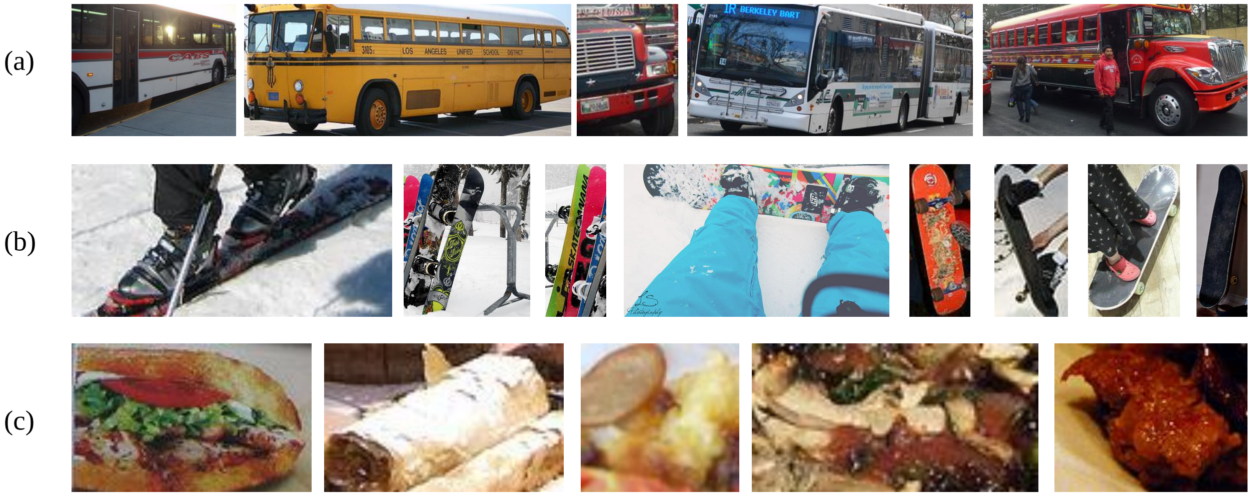

Figure 4. Visualizations of representative proposals for three background underlying categories estimated during training. They illustrate the potential conceptual overlaps between those background underlying categories and novel categories during inference on OV-COCO.

the cosine exponential score $s(\boldsymbol{w}(x), \boldsymbol{t}_{c'})$ by multiplying it with a shrinking factor, namely $1 - \sum_{c'' \in \mathcal{C}_u} P(c''|x, c')$, to alleviate this issue. Notably, the contextual embeddings of $c' \in \mathcal{C}_O$ closer to those of novel categories from $\mathcal{C}_u$ exhibit a smaller shrinking factor, demonstrating the effectiveness of our **IPR** module.

Additionally, compared to conventional designs [3, 11] on background interpretation, Figure 5 showcases more representation space diversities, conducted by the contextual embeddings of the estimated background underlying categories, further emphasizing the superiority of our LBP approach.

4. Choices of Hyper-parameters

To be consistent with VL-PLM [12], we set $\theta = 0.95$ and $\tau = 0.02$ in all experimental cases. In this section, we analyze the choices of other hyper-parameters used in the proposed approach, including n_a, γ and λ_{bg}.

Choice of n_a. To determine n_a, we compared model performance under different n_a values on OV-COCO, as outlined in Table 7. Here, $n_a = 10$ represents our default setting, acting as the baseline among its variants. When $n_a = 0$, indicating no expansion of estimated background categories, the model's performance decreased by 0.3% in AP$^n_{50}$ and 0.5% in AP$^b_{50}$ compared to the baseline. This vividly demonstrates the effectiveness of our proposed strategy to expand estimated background categories. On the other hand, increasing n_a to 20 results in a 0.1% decrease in

n_a	AP$^n_{50}$	AP$^b_{50}$	AP$_{50}$
0	37.5	58.5	53.0
10	37.8	58.7	53.2
20	37.7	58.7	53.2

Table 7. Ablation study results of our LBP approach under different n_a values on OV-COCO

γ	AP$^n_{50}$	AP$^b_{50}$	AP$_{50}$
0.01	37.6	58.8	53.2
0.02	37.8	58.7	53.2
0.05	37.5	58.7	53.1

Table 8. Ablation study results of our LBP approach under different γ values on OV-COCO.

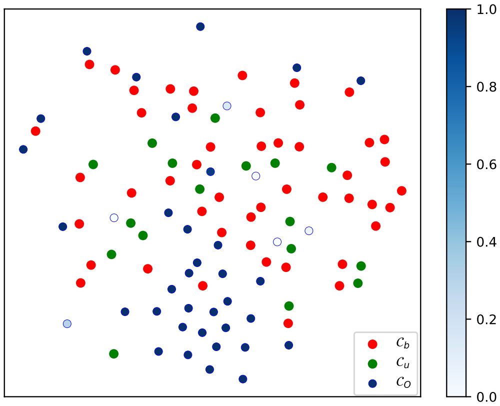

Figure 5. Visualization of the distributions of contextual embeddings of base categories $\mathcal{C}_b$, novel categories $\mathcal{C}_u$, and background underlying categories $\mathcal{C}_O$ in OV-COCO task. Here, we harness the magnitude of the shrinking factor, i.e., $(1-\sum_{c''\in\mathcal{C}_u} P(c''|x,c'))$ in Eq. (18), to showcase the semantic similarity or conceptual overlap between estimated background categories and inference novel categories. The color bar represents the relationships between $(1-\sum_{c''\in\mathcal{C}_u} P(c''|x,c'))$ for each $c'\in\mathcal{C}_O$ and the shades of blue, where darker shades indicate lower degrees of conceptual overlaps and vice versa.

λ_{bg}	AP$^n_{50}$	AP$^b_{50}$	AP$_{50}$
0.01	37.4	58.9	53.3
0.05	37.8	58.7	53.2
0.10	37.8	58.6	53.1

Table 9. Ablation study results of our LBP approach under different λ_{bg} values on OV-COCO.

performance in AP$^n_{50}$ compared to the baseline, suggesting that a larger n_a may do harm to further improve the performance of the model.

Choice of γ**.** Table 8 illustrates the detector performance of our LBP approach under different γ values. The results indicate that our method is not overly sensitive to the choice of γ. Setting γ to 0.01 or 0.05, as opposed to the default 0.02, only leads to a slight decrease in the detection performance, w.r.t. novel categories.

Choice of λ_{bg}. To better select λ_{bg}, we present its performance under different settings in Table 9. As illustrated, when $\lambda_{bg}=0.05$, the model achieves the best performance compared to that other values of λ_{bg}. Hence, we set it as the default and consider it the baseline among different variations. Specifically, reducing λ_{bg} to 0.01 resulted in a 0.4% decrease in model performance in AP$^n_{50}$ compared to the baseline. This indirectly indicates that setting a larger λ_{bg} can prevent the detector from overfitting to background underlying categories $\mathcal{C}_O$.

References

[1] 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, 2

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1

[3] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1, 2

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[8] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 1

[9] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1

[10] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021. 1

[11] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023. 1, 2

[12] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision*, pages 159–175. Springer, 2022. 1, 2