Textual Attention RPN for Open-Vocabulary Object Detection: Supplementary Material

Tae-Min Choi[1, 2]
tmchoi@kist.re.kr;tmchoi@rit.kaist.ac.kr
Inug Yoon[2]
iuyoon@rit.kaist.ac.kr
Jong-Hwan Kim[2]
johkim@rit.kaist.ac.kr
Juyoun Park[1]
juyounpark@kist.re.kr

[1] Korea Institute of Science and Technology (KIST)
Seoul, South Korea
[2] Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea

1 Training Settings

We provide more details of the training setting of TA-RPN and BARON [10] (our baseline) on COCO and LVIS datasets.

BARON settings. In BARON, several parameters are used for neighborhood sampling and classification loss. We follow the settings and network structure from BARON. For neighborhood sampling, we use the top K region proposals with an objectness score higher than 0.85, an aspect ratio between 0.25 and 4.0, and an area ratio greater than 0.01. We then apply NMS with an IOU threshold of 0.1. After filtering, we sample G bags for each filtered proposal. For COCO, we use $K = 300$ and $G = 3$. For LVIS, we use $K = 500$ and $G = 4$. For classification loss, we use $\tau = 50.0$ and $\tau = 100.0$ for COCO and LVIS, respectively.

TA-RPN settings. In TA-RPN, we use the $[eot]$ embedding from the CLIP text encoder to generate textual features. The channel numbers for visual, textual, and output features are set to $C_v = 1024$, $C_t = 512$, and $C = 1024$, respectively. The SGD optimizer with a weight decay of 0.0001 is used to train the entire network. The learning rate is initially set to 0.02 and is decreased by a factor of 10 at the 60,000th and 80,000th iterations for the COCO dataset, and at the 120,000th and 160,000th iterations for the LVIS dataset. The model is trained for 90,000 iterations on the COCO dataset and 180,000 iterations on the LVIS dataset, both with a batch size of 16. We use four RTX 3090 Ti GPUs for COCO and RTX A6000 GPUs for LVIS.

2 More Experiments

2.1 More Comparisons on COCO

In Table 1, we present additional comparison results of our method on the COCO dataset. To ensure fairness, the backbone network is standardized to ResNet50, and we clearly mark instances where extra data or a Feature Pyramid Network (FPN) is utilized. The latter enables proposal extraction from multi-resolution features. Our method outperforms the baseline, BARON, in detecting novel classes and exhibits competitive performance relative to other recently proposed methods.

Method	Backbone	Extra Data	AP_{50}^{N}	AP_{50}^{B}	AP_{50}
OVR-CNN [12]	ResNet50-C4	✓	22.8	46.0	39.9
ViLD [3]	ResNet50-FPN	✗	27.6	59.5	51.3
RegionCLIP [13]	ResNet50-C4	✓	26.8	54.8	50.4
Detic [14]	ResNet50-C4	✓	27.8	47.1	45.0
OV-DETR [11]	ResNet50-C4	✓	29.4	61.0	52.7
VLDet [7]	ResNet50-C4	✓	32.0	50.6	45.8
F-VLM [6]	ResNet50-FPN	✗	28.0	-	39.6
OADP [9]	ResNet50-C4	✗	30.0	53.3	47.2
CoDet [8]	ResNet50-C4	✗	30.6	52.3	46.6
ProxyDet [4]	ResNet50-C4	✗	30.4	52.6	46.8
BARON [10]	ResNet50-FPN	✗	<u>34.0</u>	60.4	53.5
Ours	ResNet50-C4	✗	**36.1**	53.1	43.3

Table 1: Comparison with existing OVD methods in terms of AP on the COCO dataset. We refer to the backbone structure and training source of each method. **Bold** and <u>underline</u> indicate the best and the second best performance, respectively.

Method	Backbone	Object Detection				Instance Segmentation			
		AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
ViLD [3]	ResNet50	16.1	20.0	28.3	22.5	16.3	21.2	31.6	24.4
RegionCLIP [13]	ResNet50	17.1	27.4	34.0	28.2	-	-	-	-
Detic [14]	ResNet50	-	-	-	-	17.8	26.3	31.6	26.8
DetPro [2]	ResNet50	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
OV-DETR [11]	ResNet50	-	-	-	-	17.4	25.0	32.5	26.6
OWL-ViT [1]	ResNet50	-	-	-	-	16.9	-	-	19.3
F-VLM [6]	ResNet50	-	-	-	-	18.6	24.0	26.9	24.2
Kaul et al. [5]	ResNet50	-	-	-	-	19.3	-	-	30.6
OADP [9]	ResNet50	**21.9**	28.4	32.0	28.7	<u>21.7</u>	26.3	29.0	26.6
CoDet [8]	ResNet50	-	-	-	-	**23.4**	30.0	34.6	30.7
ProxyDet [4]	ResNet50	-	-	-	-	18.9	-	-	30.1
BARON* [10]	ResNet50	20.4	30.9	33.4	30.1	19.8	28.6	30.2	27.7
Ours	ResNet50	<u>21.5</u>	30.4	33.7	30.2	20.7	28.6	30.3	27.9

Table 2: Comparison with existing OVD methods in terms of bbox AP and mask AP on the LVIS dataset with the ResNet50 backbone. * indicates the re-implementation results. **Bold** and <u>underline</u> indicate the best and the second best performance, respectively.

2.2 More Comparisons on LVIS

In Table 2, we present additional comparative results of our method on the LVIS dataset. Although our method did not outperform existing methods, it still demonstrated competitive performance. The relative underperformance can be attributed to the LVIS dataset's extensive variety, encompassing 1,203 categories. To manage memory constraints, we utilized class names from the COCO dataset as reference words. This approach, however, limited our ability to acquire sufficient textual features to generate diverse category proposals, resulting in lower performance on the LVIS dataset compared to existing methods on COCO.

2.3 More Qualitative Results

Our method not only surpasses BARON [10], our baseline, in performance on the COCO dataset, but also demonstrates this superiority through several illustrative examples, as depicted in Figure 1. The figure highlights the detection capabilities across a spectrum of categories, including both base classes (such as person, surfboard, laptop, bed, bench, chair, refrigerator, frisbee, horse, and toilet) and novel classes (including airplane, skateboard, elephant, dog, cat, sink, scissors, cake, umbrella, and snowboard). A comparative analysis of BARON and our approach reveals that our method generates more precise proposals for novel classes, and BARON occasionally struggles to detect certain novel classes. This variation in detection underscores the enhanced capability of our model to recognize and classify novel classes that were not part of its training dataset. The improvement can be attributed to the integration of rich textual features, which bolster the model's ability to interpret and respond to diverse and previously unseen objects.

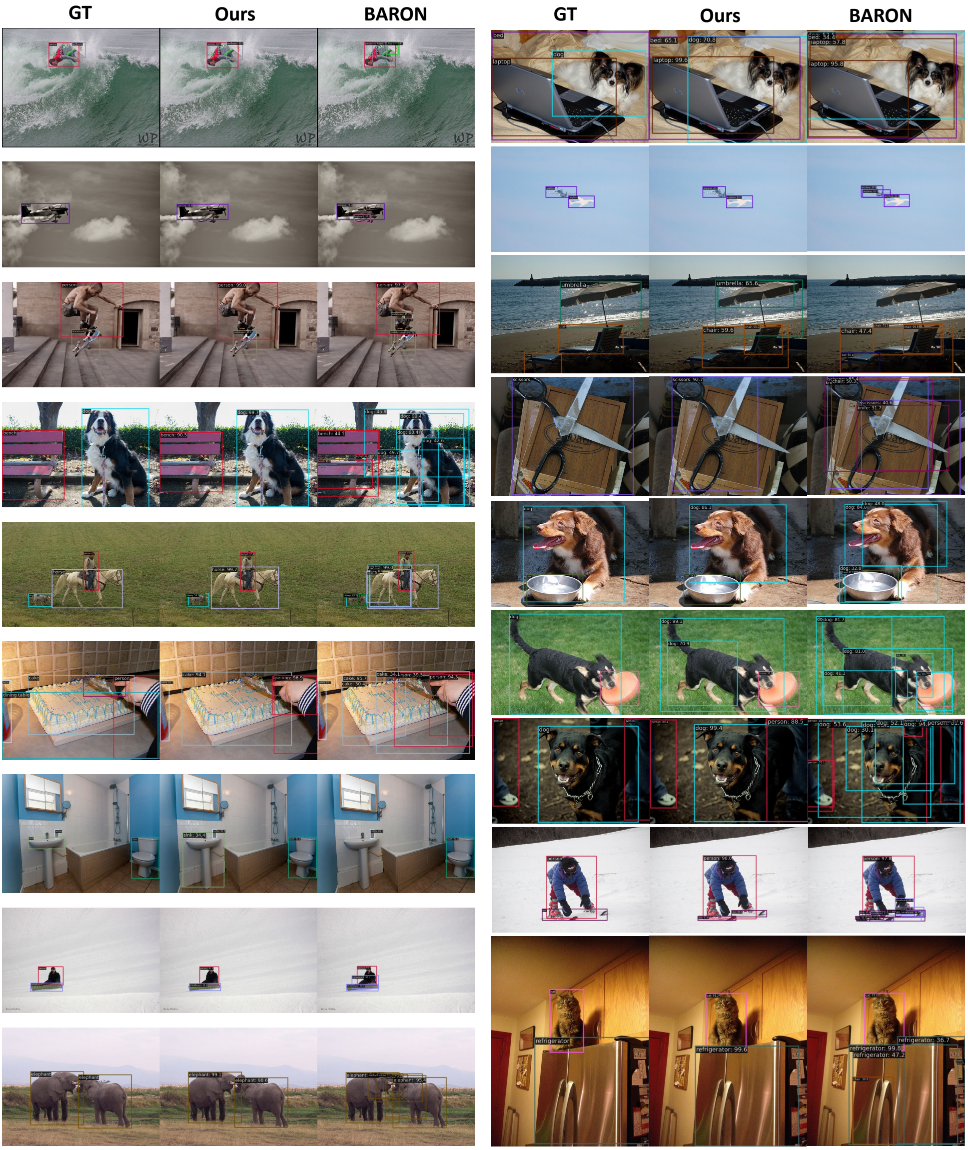

Figure 1: More qualitative results of our method and BARON on the COCO dataset.

References

[1] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022.

[2] 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.

[3] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021.

[4] Joonhyun Jeong, Geondo Park, Jayeon Yoo, Hyungsik Jung, and Heesu Kim. Proxydet: Synthesizing proxy novel classes via classwise mixup for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2462–2470, 2024.

[5] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*, pages 15946–15969. PMLR, 2023.

[6] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[7] 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. In *The Eleventh International Conference on Learning Representations*, 2022.

[8] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.