Retrieval-Augmented Open-Vocabulary Object Detection (Supplementary Materials)

Jooyeon Kim[1,*] Eulrang Cho[2,*,†] Sehyung Kim[1] Hyunwoo J. Kim[1,‡]
[1]Department of Computer Science and Engineering, Korea University [2]Samsung Research
{parang, shkim129, hyunwoojkim}@korea.ac.kr eulrang.cho@samsung.com

A. Implementation details

In this section, we provide hyperparameter settings and implementation details that are not mentioned in the main paper, for each RAL and RAF.

A.1. RAL details

After applying the rank variance sampling scheme, we set m to 2,000 and n to 10 to separate the vocabularies into hard and easy negatives. Different hyperparameters depending on the baseline and the dataset for RAL are shown in Table A1.

Method	Dataset	α^{hard}	α^{easy}	λ^{hard}	λ^{easy}	β^{hard}	β^{easy}
OADP [6]	COCO	0	1	1	5	1	1
	LVIS	0	1	1	10	1	1
Object-Centric-OVD [4]	COCO	0	1	1	10	1	1
	LVIS	0	1	1	5	1	1
DetPro [2]	LVIS	0	1	1	10	1	1

Table A1. **Hyperparameters in RAL.**

A.2. RAF details

When we bring the related concepts from the concept store, we set the number of concepts k to 50. In the augmenter $\mathcal{A}$, the number of the decoder layers L is set to 6. The one decoder layer consists of cross-attention (CA) with 8 heads and FFN with 2,048 dimensions. Positional embeddings E^{pos} and type embeddings E_1^{type} and E_2^{type} are initialized with random values. The total number of parameters for the augmenter $\mathcal{A}$ is 51M. During RAF training, we use β^{cls} of 5.0 and β^{reg} of 1.0.

B. Further ablation study

B.1. RAF in official baseline

The performance figures of the baseline compared to RALF in the main paper Section 4.2 are the results we reproduced ourselves for the fairness of the ablations. We checked the potential variation in performance depending on the execution environment. Therefore, we further verified how the performance changes when RAF is applied to the officially provided model checkpoints by the baseline authors. As depicted in Table A2, RAF demonstrates significant performance improvements on the COCO datasets.

Method	RAF	$\text{AP}_{50}^{\text{N}}$	$\text{AP}_{50}^{\text{B}}$	AP_{50}
OADP [6]		31.3	55.0	48.8
	✓	33.1	54.7	49.0
Object-Centric-OVD [4]		40.7	54.1	50.6
	✓	41.1	54.2	50.8

Table A2. **RAF ablation in the official checkpoint on COCO.**

C. Further analysis

Hyperparameters of $\mathcal{L}^{\text{RAL}}$. As depicted in Figure A1, we provide an analysis of the hyperparameters controlling $\mathcal{L}^{\text{RAL}}$ in OADP [6] on COCO dataset. To conduct the analysis, we fixed α^{hard}, λ^{hard}, and β^{hard} as (0, 1, 1) and α^{easy}, λ^{easy}, and β^{easy} as (1, 10, 1), respectively. The analysis is performed using values (0.1, 1, 10, 100) for 6 hyperparameters. When increasing α^{hard} from 1 to 10, we observed a slight improvement, however, there was a declining performance tendency overall. α^{easy} exhibited a temporary performance increase at 1 followed by a subsequent decline. Relative to lower values, λ^{easy} demonstrated better performance with higher values. Regarding λ^{hard}, β^{hard}, and β^{easy}, we found that they exhibited robust performance even with varying values.

[*] Equal contribution.
[†] This work was done when she was working at Korea University.
[‡] Corresponding author.

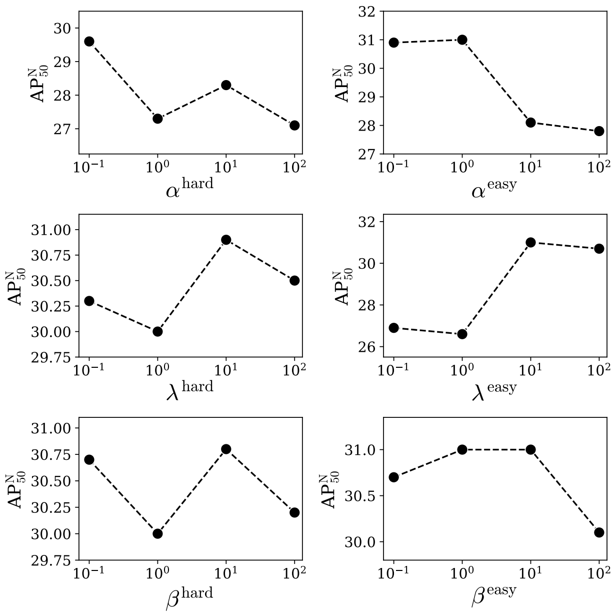

Figure A1. **Analysis on hyperparameters of $\mathcal{L}^{\text{RAL}}$.**

Exploration of n. Table A3 shows the difference between obtaining n of m negative vocabularies for each iteration based on random or similarity in RAL. The similarity performance is better than the baseline in $\text{AP}^{\text{B}}_{50}$ and AP_{50}, however, it was observed that the performance was lower in $\text{AP}^{\text{N}}_{50}$ due to the absence of randomness. Adopting randomly selecting n of m negative vocabularies showed the best performance for novel categories.

Method	$\text{AP}^{\text{N}}_{50}$	$\text{AP}^{\text{B}}_{50}$	AP_{50}
Baseline	30.0	54.5	48.1
Similarity	29.7	54.9	48.3
Random	**31.3**	54.5	48.4

Table A3. **Comparison on obtaining n.**

Sampling schemes. As described in Section 3.3 of the main paper, RAL defines hard and easy negative vocabularies through rank variance sampling of negative retrievers from the vocabulary store. To ascertain the optimality of the rank variance sampling method, we compared various sampling schemes – random sampling and similarity-based sampling, and these results are reported in Table A4. Random sampling scheme extracts vocabularies randomly without considering other factors, while similarity-based sampling scheme reflects the cosine similarity between $C^{\mathcal{V}}$ and ground-truth labels. The experimental results show that among the various sampling schemes, rank variance sampling not only demonstrates superior performance in $\text{AP}^{\text{N}}_{50}$but also exhibits outstanding performance in AP_{50}.

Sampling scheme	$\text{AP}^{\text{N}}_{50}$	$\text{AP}^{\text{B}}_{50}$	AP_{50}
Random	30.6	54.4	48.1
Similarity-based	30.2	54.8	48.3
Rank variance	**31.3**	54.5	48.4

Table A4. **Analysis on sampling scheme.**

Hyperparameter k in RAF. Table A5 and Table A6 show how much performance varies depending on the k used to augment visual features in RAF on COCO and LVIS dataset, respectively. We set k to 50 as the default. When $k = 10$, the best performance was observed on COCO. The minor performance changes are exhibited across different values of k on LVIS. The results showed that our method is robust to the choice of k.

k	$\text{AP}^{\text{N}}_{50}$	$\text{AP}^{\text{B}}_{50}$	AP_{50}
10	**33.6**	54.4	49.0
20	33.3	54.4	48.9
50	33.4	54.5	49.0
100	33.3	54.4	48.9

Table A5. **Analysis on k in RAF on COCO.**

k	AP_{r}	AP_{c}	AP_{f}	AP
10	**21.9**	26.2	29.1	26.6
20	**21.9**	26.2	29.1	26.6
50	**21.9**	26.2	29.1	26.6
100	21.8	26.2	29.1	26.6

Table A6. **Analysis on k in RAF on LVIS.**

Scale of large vocabulary. As discussed in Section 3.3 of the main paper, RALF uses a large vocabulary set to construct the vocabulary store. We set the experiment to analyze the effectiveness of the vocabulary size. We adopt V3Det [5] with 13,204 vocabularies as the vocabulary set for all experiments. Before retrieving hard and easy negatives from the vocabulary store, we construct the vocabulary store by eliminating unnecessary elements, *i.e.*, novel categories and case-overlapping, from the vocabulary set. Following this process, we obtained 13,064 vocabularies from V3Det. To examine whether vocabulary size affects performance, we experiment on COCO benchmarks by reducing the percentage of vocabulary size to 40% and 70%. The results are shown in Table A7. From the results, we observed that the performance was better when the vocabulary size was 100% compared to when it was 40% or 70%.

Retrieve negative vocabularies with BERT. In this work, the hard and easy negative vocabularies are retrieved based

Scale of large vocabulary	AP^N_{50}	AP^B_{50}	AP_{50}
40%	30.9	54.7	48.5
70%	30.8	54.7	48.4
100%	**31.3**	54.5	48.4

Table A7. **Analysis on a large vocabulary size.**

on cosine similarity between CLIP [3] text embeddings in RAF. When retrieving negative vocabularies, it is also possible to use the embeddings of a language model (LM) instead of CLIP. We extract embeddings about base categories and a large vocabulary set with a language model BERT [1] and then retrieve hard and easy negative vocabularies based on the cosine similarity between the embeddings. Table A8 depicts the comparison results of CLIP and BERT. The performance on the novel categories is lower in BERT, but it increases by 0.7 AP^N_{50} compared to the baseline.

Method	AP^N_{50}	AP^B_{50}	AP_{50}
Baseline	30.0	54.5	48.1
CLIP [3]	**31.3**	54.5	48.4
BERT [1]	30.7	55.4	48.9

Table A8. **CLIP vs BERT.**

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 3

[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 *CVPR*, 2022. 1

[3] 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 *ICML*, 2021. 3

[4] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*, 2022. 1

[5] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *ICCV*, 2023. 2

[6] 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 *CVPR*, 2023. 1