CoDet: Co-Occurrence Guided Region-Word Alignment for Open-Vocabulary Object Detection

Anonymous Author(s)
Affiliation
Address
`email`

Contents

A A Heuristic Baseline for Co-occurrence Discovery

In this section, we introduce the baseline method used for ablation study in Table 4b (main paper) in more detail. This baseline is adapted from a recently proposed image co-segmentation method ReCo [6]. As shown in Figure 1, it basically consists of four steps to identify the co-occurring object in the query image: First, it estimates pair-wise region similarity between region proposals of the query image and support images, which is the same as CoDet. This yields a similarity matrix $\mathbf{S} \in \mathbb{R}^{n \times m \times n}$, where n stands for the number of proposals per image, and m stands for the number of support images. Second, it applies a $\max$ operator on the last dimension of $\mathbf{S}$, which serves to find the nearest neighbor in each support image for each region proposal in the query image. This reduces $\mathbf{S}$ to an $n \times m$ matrix. Third, it applies a mean operator on the second dimension of $\mathbf{S}$ to derive the average support that each proposal has among the support images. Finally, it identifies the co-occurring object as the one with the highest average maximum similarity (support) among support images, by applying an argmax operator on the first dimension of $\mathbf{S}$.

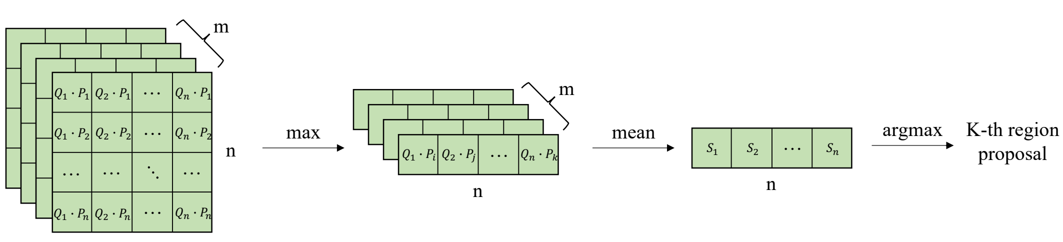

Figure 1: **Illustration of the baseline method for co-occurrence discovery**. Q and P are region proposals in the query image and support images, respectively. S is the averaged maximum similarity score across support images.

B Further Analysis on Different Alignment Strategies

Complementing the discourse in Section 4.5 (main paper), we further delineate the performance of different alignment strategies with respect to novel category AP_{50} on OV-COCO in Figure 2. It can be seen that strategies based on region-region similarity or hand-crafted rules (max-size) show steady improvement in novel object recognition across training, whereas the performance of region-word similarity-based method is highly unstable and even decreases in the early stage. A possible explanation is that solely relying on region-word similarity to align regions and words may be more susceptible to errors in pseudo-labels. For instance, if the model incorrectly matches the text label 'seagull' with the object 'dove' at the initial phase, its supervision signal would pull the two closer in the shared feature space. This negative feedback could directly harm the following pseudo-labeling process, thus, there is a higher probability for the model to make the same mistake.

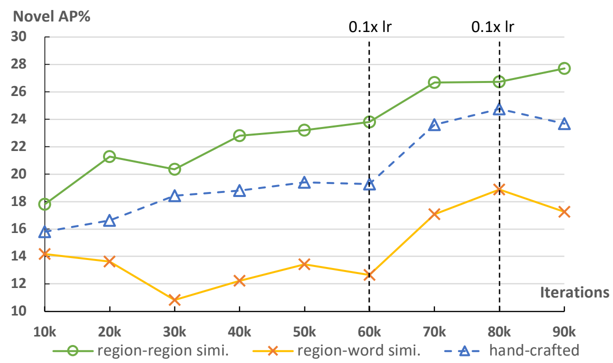

Figure 2: Performance of different alignment strategies at discrete training stages on OV-COCO.

C Visualization on OV-LVIS and OV-COCO

We visualize more detection results of CoDet in Figure 3 and Figure 4. On OV-LVIS, we can see that CoDet successfully detects many rare objects, *e.g.*, gas mask, puffin, horse buggy, heron, satchel, and so on (Figure 3). This validates that CoDet can efficiently leverage web-crawled image-text pairs to learn open-word knowledge for novel object recognition. On OV-COCO, our method continues to demonstrate strong open-vocabulary capability and correctly detects some hard samples, *e.g.*, the occluded 'tie' and 'elephant' (upper left of Figure 4). Nevertheless, we also notice that the prediction scores for novel categories are generally lower than base categories, which suggests the model is biased towards base classes in OV-COCO. Such tendency to overfit base categories is also observed in other works [8, 3, 7], due to the small training vocabulary of OV-COCO. We believe adopting tricks like focal loss could alleviate this issue and further benefit our method.

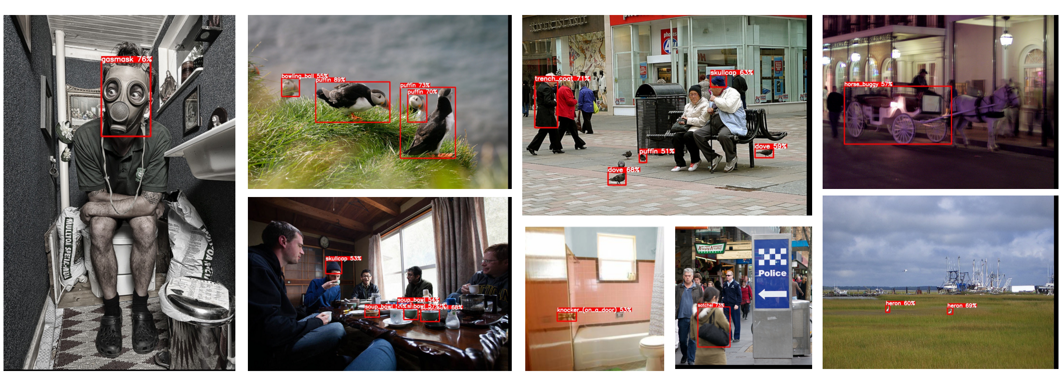

Figure 3: **Visualization of prediction results by CoDet on OV-LVIS**. For clarity, we only show results for novel categories.

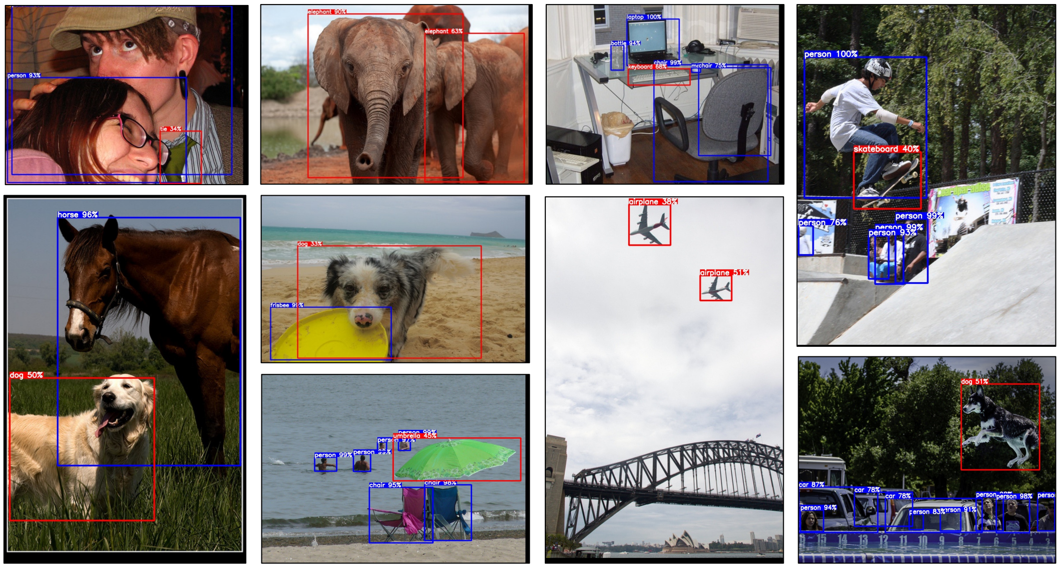

Figure 4: **Visualization of prediction results by CoDet on OV-COCO**. Red boxes are for novel categories, while blue boxes are for base categories.

D Implementation Details

Table 1 lists the detailed hyper-parameter configuration used for our OV-LVIS and OV-COCO experiments. We follow Detic [9] to use low input resolution and large batch size for caption data to achieve better trade-off between efficiency and performance. For experiments employing Swin-Base [5] as the visual backbone, the settings are mostly the same as ResNet50 [2], except that a higher resolution (896 for detection and 448 for caption) is adopted to maintain fair comparison with [9, 4].

Table 1: **Hyper-parameter configuration of CoDet.** LSJ stands for large scale jittering [1]. Resolution refers to the resized short side length of input images.

Configuration	**OV-LVIS**	**OV-COCO**
Optimizer	AdamW	SGD
Gradient clipping	True	True
Learning rate (LR)	2e-4	2e-2
Total iterations	90k	90k
Warmup iterations	1k	–
Step decay factor	–	0.1×
Step decay schedule	–	[60k, 80k]
Data augmentation	LSJ	none
Batch size (detection)	8	2
Batch size (caption)	32	8
Resolution (detection)	640	800
Resolution (caption)	320	400
Detection/Caption data ratio	1:4	1:4
Federated loss [10]	True	False
Repeat factor sampling	True	False
$\mathcal{L}_{\text{region-word}}$ weight	0.2	0.1
$\mathcal{L}_{\text{image-text}}$ weight	0.2	0.1

References

[1] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021. 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[3] 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*, 2023. 2

[4] 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*, 2023. 3

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[6] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. In *Advances in Neural Information Processing Systems*, volume 35, pages 33754–33767, 2022. 1

[7] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. *arXiv preprint arXiv:2302.13996*, 2023. 2

[8] 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. 2

[9] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022. 3

[10] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 3