Open-Vocabulary DETR with Conditional Matching

Yuhang Zang[1], Wei Li[1], Kaiyang Zhou[1], Chen Huang[2], Chen Change Loy[1]✉
[1]S-Lab, Nanyang Technological University [2]Carnegie Mellon University
{zang0012, wei.l, kaiyang.zhou, ccloy}@ntu.edu.sg chen-huang@apple.com

Supplementary Material

A More Qualitative Results of OV-DETR

Open-Vocabulary COCO. We provide more qualitative results of OV-DETR on Open-Vocabulary COCO setting (Fig. 1). We visualize the detection results on *novel* classes and the activation maps of OV-DETR and Region Proposal Network (RPN) [3] used by ViLD [1], which further validate the motivation from the main paper: OV-DETR has higher activation values on objects of *novel* classes than RPN.

Web Images. To verify the generalization ability, we also provide the qualitative results of anime characters in Fig. 2. Although these characters are not provided during training, OV-DETR can successfully detect the regions matched with the conditional image queries.

Failure Cases. Fig. 3 shows some failure cases of OV-DETR. We notice that detecting small or occluded objects with conditional image query is hard. Our method is not robust to the unrelated out-of-distribution text queries. We will address these shortcomings in further research.

B Discussion of Object Proposals

In previous work ViLD [1], class-agnostic object proposals are leveraged to transfer the knowledge from CLIP image encoder to the detector. As shown in Fig. 4 (a), ViLD first trains a RPN network on *base* classes to get M pre-computed proposals. These object proposals may contain objects of *novel* classes and are essential for training ViLD model. For these M proposals, M predicted region embeddings and the corresponding ground-truth embeddings are computed by a Mask R-CNN detector and a CLIP image encoder respectively. Then, a knowledge distillation loss is applied for the predicted region embeddings and the ground-truth embeddings.

We follow the steps of ViLD, pre-training a detector with the *base* classes to predict object proposals that may cover the *novel* classes. The only difference is that we use different architectures (Def-DETR vs. the RPN network in ViLD). Despite of building upon different architectures, the generated object proposals have similar high top-300 averaged recall (AR@300) for *novel* categories (48.3 for ViLD and 47.6 for ours). Fig. 4 (b) shows that compared with ViLD, we

Table 1: Importance of $\mathcal{L}_{\text{embed}}$ on LVIS.

#	$\mathcal{L}_{\text{embed}}$	AP^{m}	$\text{AP}^{\text{m}}_{\text{novel}}$	$\text{AP}^{\text{m}}_{\text{c}}$	$\text{AP}^{\text{m}}_{\text{f}}$
1	✗	24.9	14.4	23.2	31.3
2	✓	**26.6**	**17.4**	**25.0**	**32.5**

Table 2: **Ablation study** on N (the number of object queries) and R (the number of copies).

#	N	R	AP^{m}	$\text{AP}^{\text{m}}_{\text{novel}}$	$\text{AP}^{\text{m}}_{\text{c}}$	$\text{AP}^{\text{m}}_{\text{f}}$
1	100	1	22.0	10.6	20.9	28.2
2	100	3	25.7	13.6	**25.0**	31.9
3	100	9	24.3	11.9	22.9	31.3
4	300	1	24.2	12.3	22.8	30.9
5	300	3	**26.6**	**17.4**	**25.0**	**32.5**

use the object proposals in a different way. Since these object proposals are class-agnostic, they cannot be applied on DETR's matching algorithm. In OV-DETR, we treat the image embeddings extracted by the object proposals as the conditional image query. The expected predictions of OV-DETR are the 'matched' regions of an input image given a conditional image query.

C Importance of $\mathcal{L}_{embed}$

In OV-DETR, we introduce an embedding reconstruction head to predict the conditional input embedding $\boldsymbol{z}^{\text{text}}$ or $\boldsymbol{z}^{\text{image}}$, and this reconstruction head is optimized by the loss $\mathcal{L}_{embed}$. The results in Table 1 show the efficacy of $\mathcal{L}_{embed}$.

D Importance of Multiple Queries for Training

Recap that we propose to "clone" query features in Section 3.2 (also see Fig 4) of the main paper. We examine different choices of the two hyper-parameters N and R, and show results in Table 2. When $N = 100$, we find that coping queries (from $R = 1$ to 3) improves the $\text{AP}^{\text{m}}_{\text{novel}}$ from 10.6 to 13.6, and a slight degradation when $R = 9$ partially due to the limited optimization capacity. When $N = 300$, we observe that coping queries (from $R = 1$ to 3) is also beneficial. However, we will face the out-of-memory issue on GPU when $N = 300$ and $R > 3$. Overall, we find the combination of $N = 300$ queries and repetition of $R = 3$ times serves as the optimal solution.

E More Implementation Details

Hyper-Parameters. All models are trained on 8 Tesla V100 GPUs. We use the ResNet50-C4 backbone as our default choice. We keep most of the hyper-parameters the same with previous works [1,7]. For loss functions, we set the weighting parameters $\mathcal{L}_{\text{BCE}} = 3.0$, $\mathcal{L}_{\text{L1}} = 5.0$, $\mathcal{L}_{\text{GIoU}} = 2.0$ and $\mathcal{L}_{\text{embed}} = 1.0$. The input resolution of the CLIP model is set to 224x224, and the temperature τ of the CLIP model is set to 0.01.

Text Prompts. Prompt tuning is a critical step when transferring pre-trained language models to downstream computer vision tasks [6,5,4]. We follow the same process as in ViLD [1] to construct the text prompts. Specifically, for each class we feed the textual name wrapped in 63 different prompt templates (e.g., `‘there is a {class name} in the photo’`) to CLIP’s text encoder, and then average the 63 text embeddings, which is known as prompt ensembling [2].

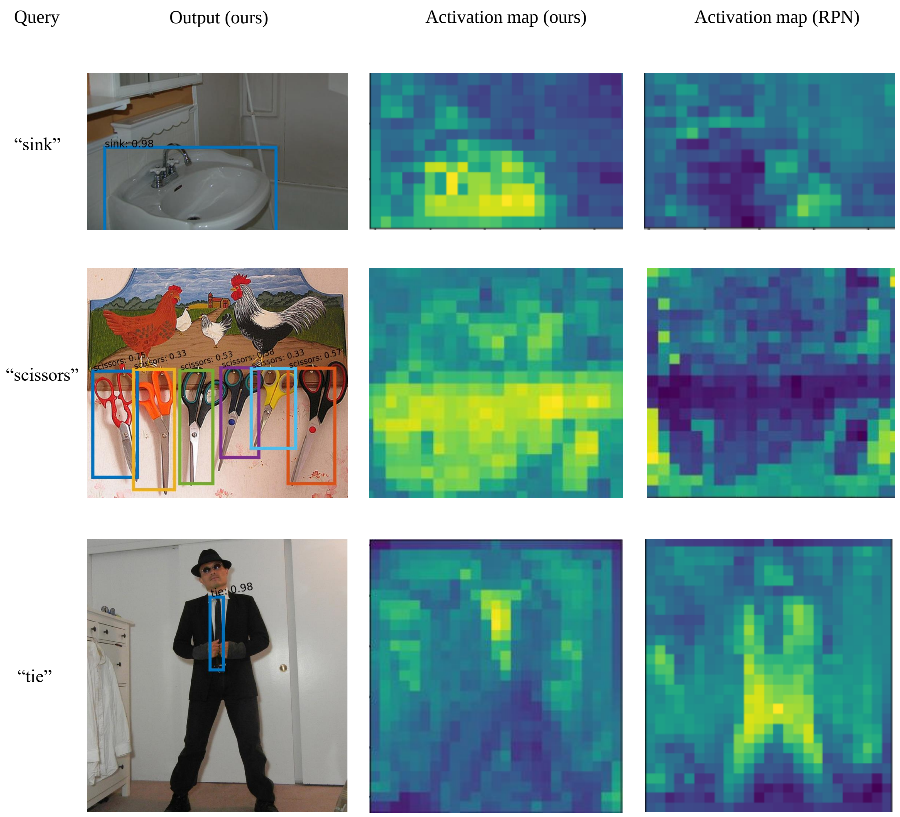


Fig. 1: **Qualitative results on Open-Vocabulary COCO setting.** We visualize the prediction results of OV-DETR on *novel* classes. We also provide the comparison of activation maps between ours and the RPN network.

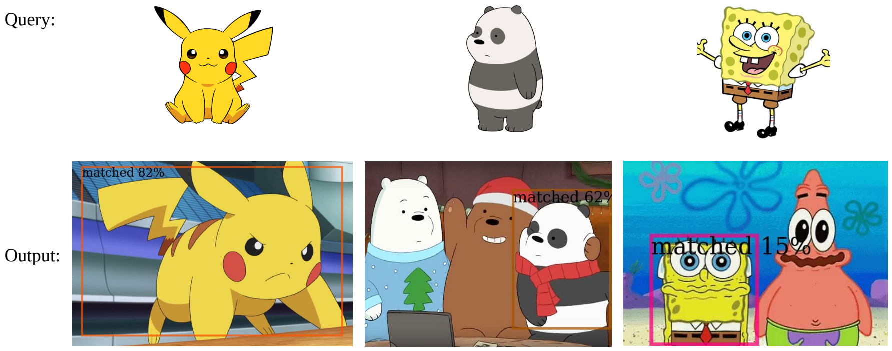


Fig. 2: **Qualitative results on anime characters.** These images are collected from web. We use the model trained on LVIS dataset to check the matchability with the given conditional image queries.

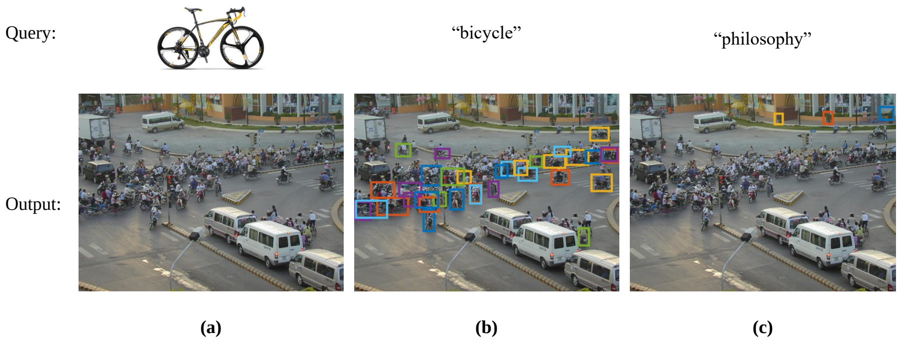
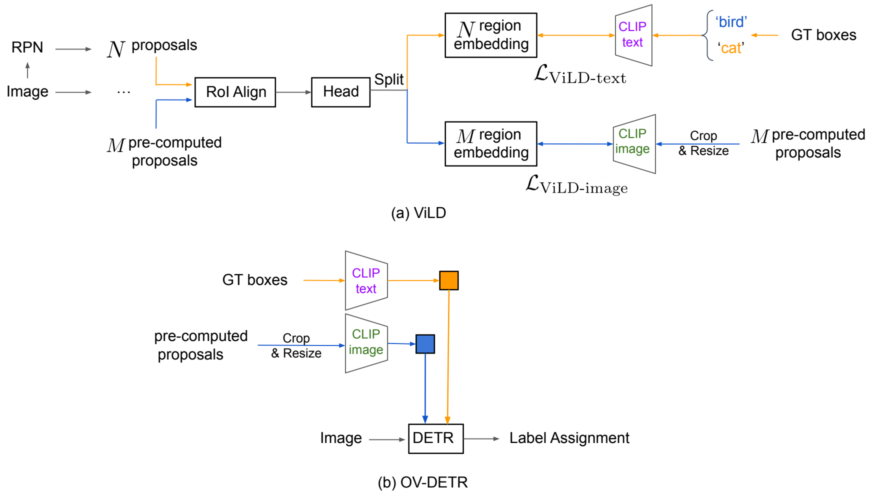


Fig. 3: **Failure cases of OV-DETR.** These images are collected from the web. We use the model trained on LVIS dataset to check the matchability with the given conditional image or text queries. **(a):** OV-DETR fails to detect these small and occluded objects with the conditional image query ("bicycle"). But as shown in **(b)**, this issue can be solved to some extent by using text query. **(c):** Given the unrelated text queries (*e.g.*, "philosophy"), OV-DETR will predict wrong false-positive detection results.



(a) ViLD

(b) OV-DETR

Fig. 4: The comparison of ViLD and our OV-DETR of utilizing pre-computed object proposals. **(a):** ViLD leverages a distillation loss between the predicted region embeddings and pre-computed object proposals. **(b):** We use pre-computed object proposals to generate the conditional image query.

References

1. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2022)
2. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS **28**, 91–99 (2015)
4. Zhang, Y., Zhou, K., Liu, Z.: Neural prompt search (2022)
5. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022)
6. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV (2022)
7. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2020)