Supplementary Materials
NoOVD: Novel Category Discovery and Embedding for Open-Vocabulary Object Detection

Yupeng Zhang[12] Ruize Han[3] Zhiwei Chen[4] Wei Feng[12] Liang Wan[12*]
[1]College of Intelligence and Computing, Tianjin University.
[2]Key Research Center for Surface Monitoring and Analysis of Relics, State Administration of Cultural Heritage.
[3]Faculty of Computer Science and Artificial Intelligence, Shenzhen University of Advanced Technology.
[4]School of Artificial Intelligence, Nanchang University.
{zhangyupeng, wfeng, lwan}@tju.edu.cn, hanruize@suat-sz.edu.cn, zhiweichen@ncu.edu.cn

1. Test Stage

As shown in Fig. 1, during the test stage, we also adopt the latent novel-category object discovery method proposed in Section *Novel-Category Embedding by Distillation* to filter RPN-generated proposals and identify regions likely to contain novel-category objects. Specifically, we first utilize category-agnostic foreground-background text descriptions generated by ChatGPT-o1 and extract their embeddings using the frozen CLIP text encoder. We then compute the cosine similarity between these embeddings and the proposal features to estimate the likelihood of each proposal containing a novel-category object. The selected proposals are subsequently passed through the R-RPN for confidence rectification and sorted based on the adjusted scores. Finally, following the standard RPN post-processing pipeline, we select the top 1000 proposals and feed them into the RoI head for bounding box regression and classification, thereby improving the recall of novel-category objects.

2. Inference Efficiency Analysis

We test the inference time on a single 3090 GPU. The original F-ViT + DeCLIP (ViT-B/16) achieves FPS of **21.18**, while our NoOVD + DeCLIP (ViT-B/16) achieve FPS of **14.11**. While K-FPN and R-RPN introduce a certain level of overhead during inference, reducing FPS from 21.18 to 14.11, this trade-off yields a 2.9% gain in AP_r — an acceptable balance in most open-vocabulary scenarios. We note that existing methods, such as DeCLIP, often omit reporting inference efficiency. To address this gap, we benchmarked two representative approaches: MM-OVOD [1], and OV-DQUO [2], which achieve FPS of **2.09** and **2.31**, respectively. In contrast, our method not only maintains higher accuracy but also achieves significantly better runtime performance.

Table 1. Recall of K-FPN and RPN.

Method	Dataset	$AR_{\text{OV-COCO}}$	$AR_{\text{OV-LVIS}}$
RPN	Train set	61.1%	45.8%
K-FPN	Train set	**79.5%**	**73.6%**
RPN	Test set	64.9%	36.3%
K-FPN	Test set	**82.6%**	**67.5%**

3. Visualized Results

As shown in Fig. 2, we present the visualization results of foreground object discovery by DeCLIP ViT-B/16. Guided by the foreground–background text prompts generated by the LLM, the proposed K-FPN effectively highlights the salient foreground regions within an image. Although some proposals exhibit spatial overlap or slight positional deviations, this is expected at this stage, as these regions are still initial candidates that have not yet undergone refined regression by the RoI head. Since the distillation process operates at a coarse level, it is inherently robust to minor localization errors, and these deviations do not affect subsequent training. **This also further demonstrates that K-FPN preserves and leverages the pretrained semantic knowledge of CLIP effectively.** To assess the coverage of K-FPN on rare-category ground-truth objects, we use *Recall*@IoU$\geq$0.5 as the evaluation metric. As shown in Table 1, K-FPN achieves substantially higher recall than the original RPN on both the training and test sets of OV-COCO and OV-LVIS. These results verify that K-FPN is able to identify a large number of potential novel-category regions, thereby providing crucial and high-quality supervisory signals for subsequent knowledge self-distillation.

In Fig. 3, we compare the detection results of GT, F-ViT + DeCLIP, and our NoOVD + DeCLIP. The results clearly demonstrate the superior robustness of NoOVD, particu-

*Corresponding author.

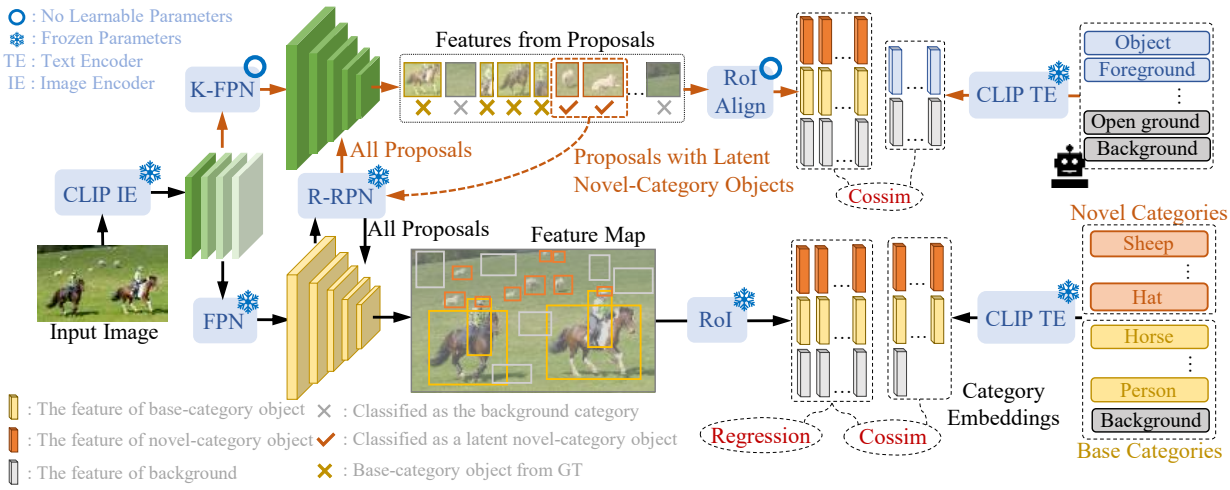

Figure 1. Illustration of the testing process of NoOVD.

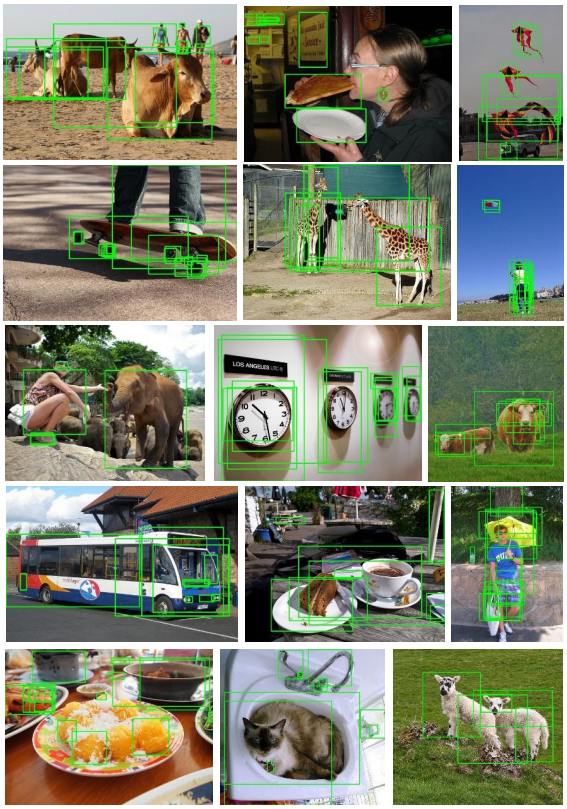

Figure 2. Examples of foreground objects detected through K-FPN and text descriptions generated by LLM.

larly in detecting novel-category objects. For instance, in images with red bounding boxes, our method successfully identifies previously novel-category objects such as 'mini candy', 'matchbox', and 'garbage', which are missed by F-ViT + DeCLIP. In addition, our bounding boxes exhibit higher localization accuracy compared to F-ViT + DeCLIP in many cases. Notably, NoOVD also detects additional plausible objects that are not annotated in the GT, suggesting its enhanced generalization capability. Overall, NoOVD delivers more comprehensive and accurate object detection, excelling particularly in novel-category objects, and offers new insights for the advancement of OVD.

References

[1] 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. 1

[2] Junjie Wang, Bin Chen, Bin Kang, Yulin Li, Weizhi Xian, Yichi Chen, and Yong Xu. Ov-dquo: Open-vocabulary detr with denoising text query training and open-world unknown objects supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7762–7770, 2025. 1

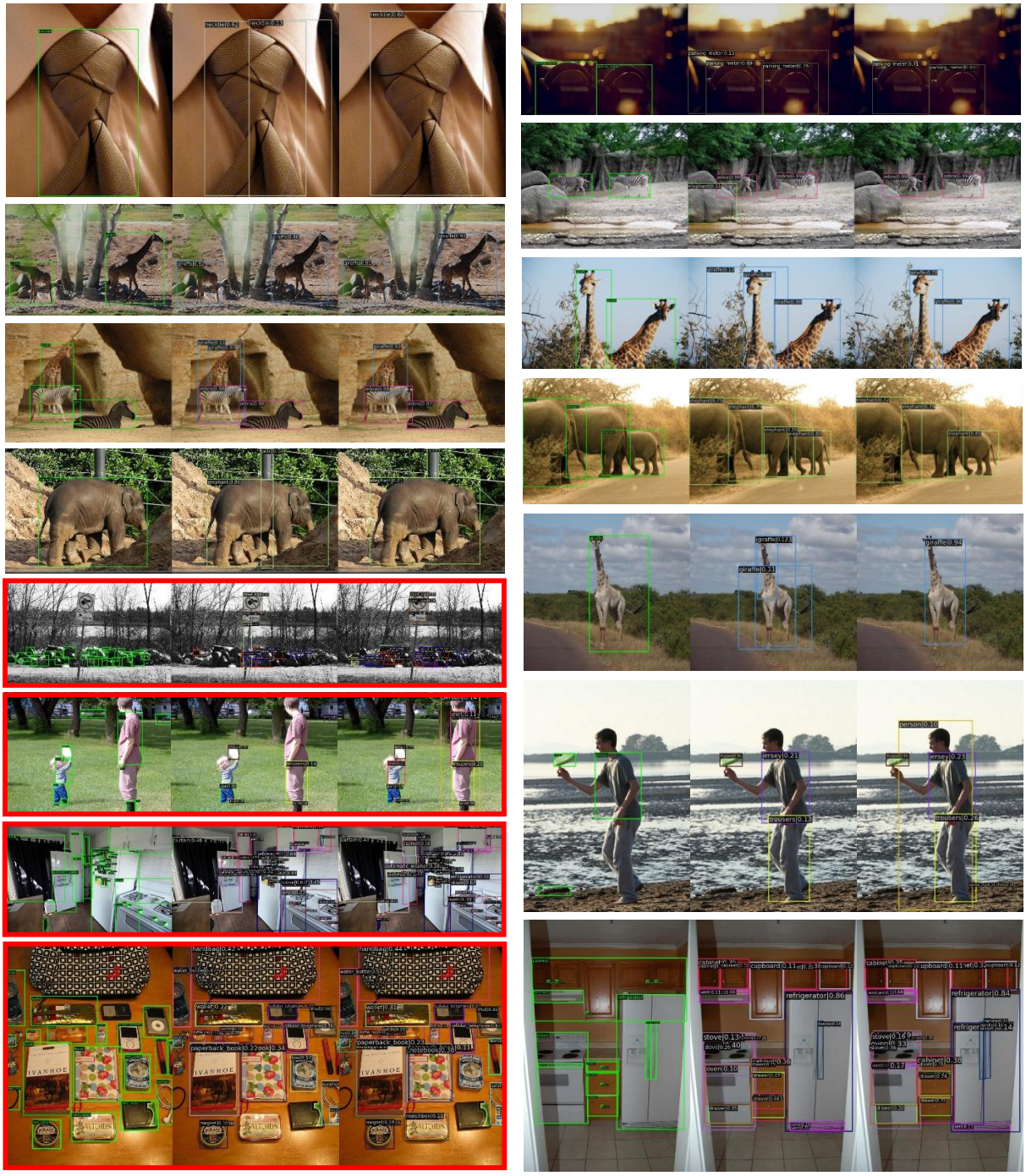

Figure 3. The visualized results of GT, F-ViT+DeCLIP and NoOVD+DeCLIP.