Supplementary Materials
ODOV: Benchmark the Open-Domain Open-Vocabulary Object Detection

Yupeng Zhang[12] Ruize Han[3*] Fangnan Zhou[1] 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.
{zhangyupeng, zhoufangnan, wfeng, lwan}@tju.edu.cn, hanruize@suat-sz.edu.cn

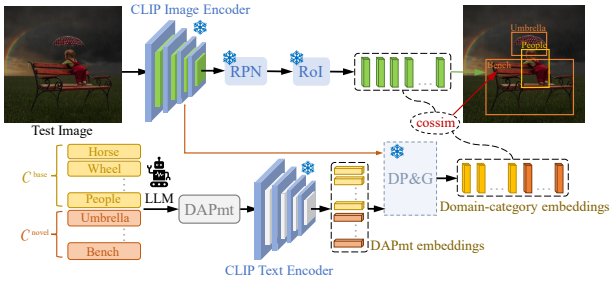

Figure 1. Overview of DVtor testing framework. The method dynamically generates fused embeddings based on the domain characteristics of each test image, enabling robust recognition of novel categories across diverse domains.

1. Testing Stage

As shown in Fig. 1, during testing, the images come from diverse open domains so we do not apply the random perturbations on the features for simulative domain argumentation. Similar to the training phase, we use the trained DP&G to extract domain information to generate domain embeddings, which is fused with the DAPmt embeddings for both *base and novel* categories. Finally, the RPN-extracted proposals are used for bounding box localization, and cosine similarity with the domain-category fused embeddings is calculated for object classification.

2. More Details of OD-LVIS

2.1. Data Generation, Cleaning and Annotation

Open-domain generation for various imaging conditions. We generate images with six types of imaging conditions from clear images to enhance the domain diversity. The specific methods are as follows: ① *Rainy Images*: Leveraging [11], we created diverse rainy scene images to enrich the dataset's representation of rain scenarios. ② *Hazy Images*: Following [12], we synthesized hazy images with varying concentrations by altering the atmospheric scattering coefficient. ③ *Illumination Variations*: Illumination conditions were diversified using gamma correction, with different gamma values applied to adjust the lighting. ④ *Low Resolution*: Based on [1, 5], high-resolution images were downsampled using a bicubic kernel, with the degradation level controlled by the downsampling factor. ⑤ *Noise*: Gaussian White Noise: Inspired by [6], we synthesized Gaussian white noise images with varying degradation levels by modifying the noise variance. Salt-and-Pepper Noise: Similarly, we added salt-and-pepper noise with degradation intensity controlled by noise density [3]. ⑥ *Blur*: Gaussian Blur: Following [3], we added Gaussian blur with degradation intensity controlled by Gaussian kernel standard deviation. Motion Blur: Following [8], we synthesized images with varying degrees of motion blur by adjusting the blur kernel length. Out-of-focus: Using the method from [7], we generated defocused images by adjusting the radius of a circular averaging filter.

Open-domain generation for various image styles. In addition, we generate nine distinct image styles to further enhance the diversity of the dataset. Specifically, we employ Baidu's Style Transfer API to synthesize various styles, including pencil sketches, watercolors, *etc*. These styles are widely prevalent on the Internet, particularly with the rapid advancement of generative artificial intelligence. Such stylized images provide more diverse and challenging data for open-domain object detection.

Cleaning and annotation. After completing data generation, we establish strict selection criteria to ensure the benchmark's quality and reliability. To guarantee the robustness of this process, we invite multiple researchers in related fields to participate in and oversee the selection. Manual inspection ensures that images and objects remain clear and intact, with categories easily recognizable. Since the data generation is based on LVIS [4], we preserve the original annotations in full, including both object detection

*Corresponding author.

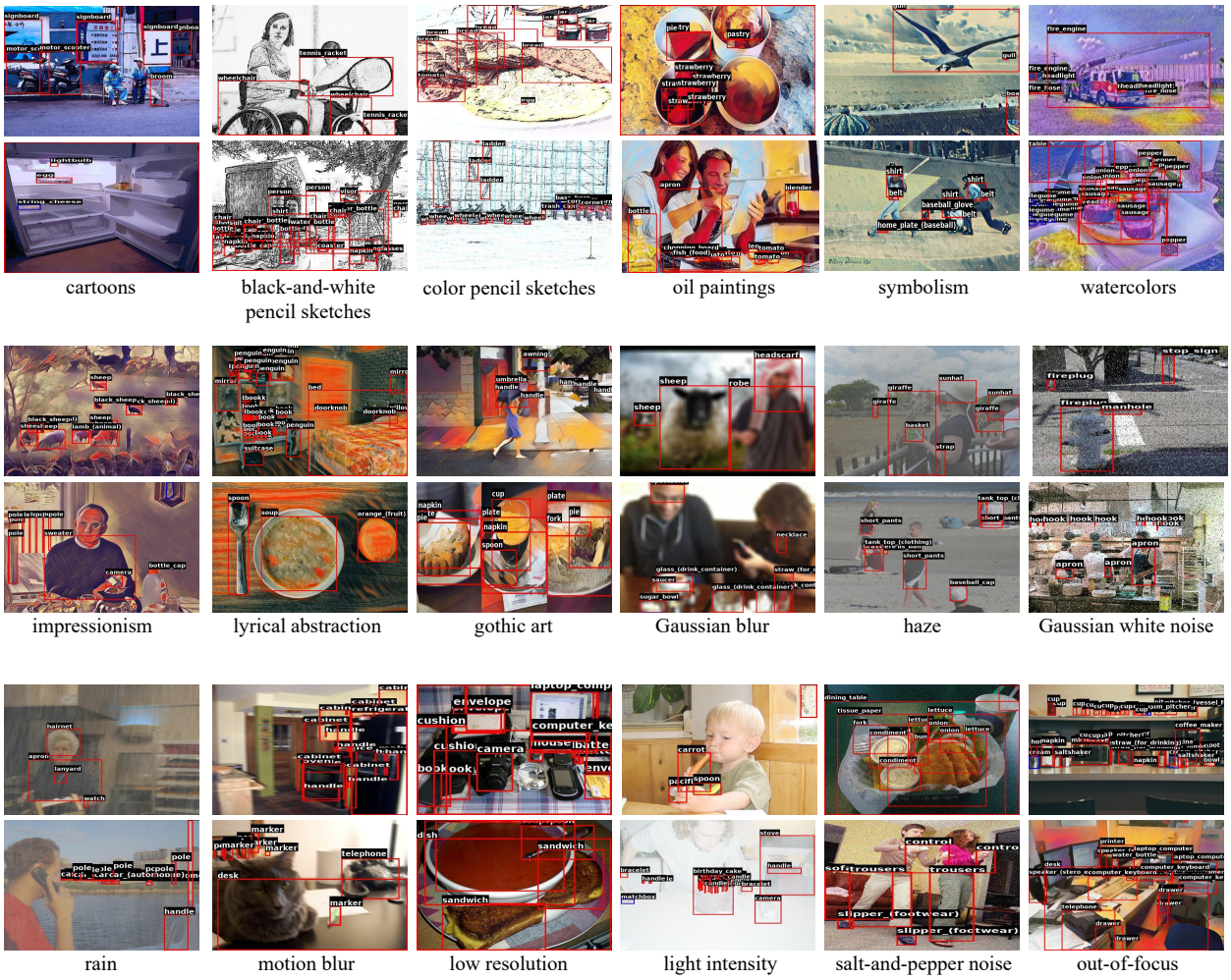

Figure 2. Some examples from OD-LVIS, covering all domains in the dataset (including three types of blur and two types of noise), along with their annotations.

and instance segmentation labels. Importantly, the domain expansion of OD-LVIS does not alter the original annotations, as no scaling or deformation of objects occurs during synthesis. We then organize the annotations and standardize them into both Pascal VOC [2] and LVIS formats.

2.2. Characteristics

Figure 2 shows several examples from OD-LVIS, covering all domains (two types of noise and three types of blur) along with their annotations. We then present the characteristics of the proposed OD-LVIS, which contains a series of challenging settings specifically designed to evaluate the limitations of object detection models for open-world scenarios.

Multi-category and multi-object scenes. Most images in OD-LVIS contain multiple objects from different categories, evaluating the model's robustness in both object localization and classification.

Various object sizes and aspect ratios. OD-LVIS encompasses the same category of objects but with different sizes and shapes, requiring models to recognize and distinguish objects under diverse visual conditions.

Complex backgrounds and overlapping objects. Lots of images in OD-LVIS contain complex backgrounds and overlapping objects, reflecting the wild environments surrounding objects in real-world scenarios.

Long-tailed category distribution. Similar with the object categories in real world, which in our dataset also follow a long-tailed distribution, demanding models to exhibit strong capabilities in localizing and distinguishing rare or tail-class samples.

Cross-domain variability. Since OD-LVIS shares categories with LVIS, it can be combined with LVIS (used for training) to further assess the domain generalization ability

Figure 3. Visualization of open-domain open-vocabulary object detection results (under various image styles).

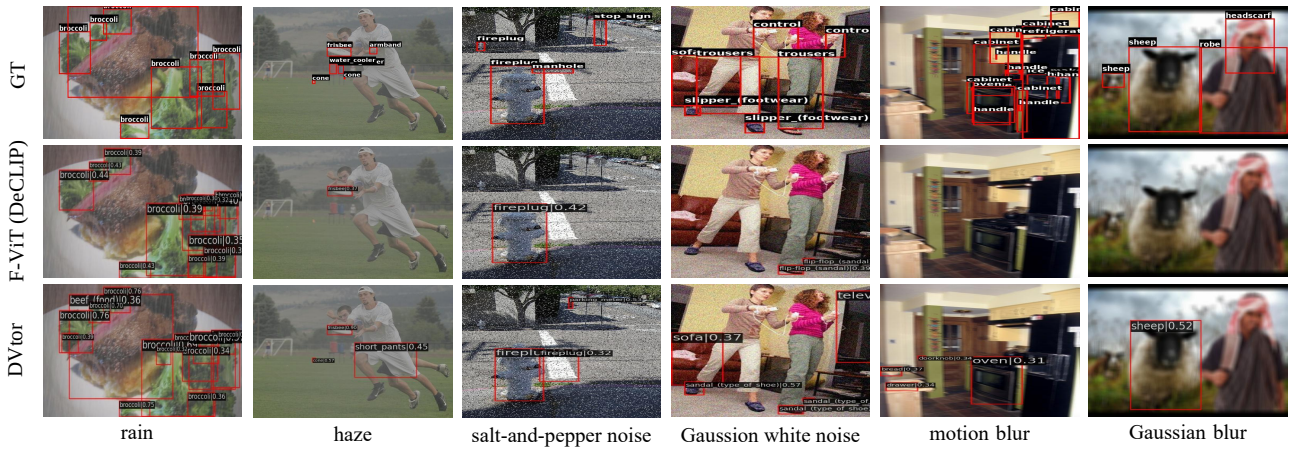

Figure 4. Visualization of open-domain open-vocabulary object detection results (under various imaging conditions).

of object detection models across 15 various scenes.

These challenges are designed to simulate complex real-world conditions, providing a comprehensive benchmark to advance the development of detection technologies in open real-world.

3. Visualization Result Analysis

We present the qualitative results of ODOV object detection on OD-LVIS using both F-ViT [10] (DeCLIP [9]) and DVtor. These visualizations highlight the models' performance and their ability to detect novel objects. Notably, in certain complex scenarios, F-ViT (DeCLIP) struggles to effectively detect objects, whereas DVtor demonstrates superior adaptability, accurately identifying the majority of objects in more challenging environments.

As shown in Fig. 3, we can see that, with the image style shift, many targets (such as the 'car' in the first column, and the 'baseball bat' in the second column) can not be detected by the F-ViT (DeCLIP) as shown in the second row. The proposed method can effectively detect these targets. More seriously, F-ViT (DeCLIP) can not identify any object under the more abstract styles, such as black-and-white pencil sketches, lyrical abstraction, and oil paintings as shown in the last three columns. The proposed method can identify many of the interested objects, which demonstrates our approach's ability to detect the objects in complex and dynamic scenarios compared to previous methods. However, compared to the ground-truth results, DVtor also fails to detect many objects. This also indicates that the proposed ODOV is very difficult with much room for improvement.

As shown in Fig. 4, with the imaging condition shift, F-ViT (DeCLIP) performs well in some cases, such as the (light) rainy scene in the first column. It is partly influenced in the following cases, including the haze, salt-and-pepper noise, and Gaussian white noise, in which DVtor performs better to correctly detect more targets. As shown in the last

Figure 5. Illustration of the failure cases.

two columns, the motion blur and Gaussian blur are severe, which makes the F-ViT (DeCLIP) difficult to detect the desired objects. Under these cases, DVtor can also identify some of them, which further demonstrates the effectiveness of DVtor.

Failure cases. As shown in Fig. 5, we present typical false detections (*e.g.*, horse carriage → bicycle, dog → cat) and missed detections in cross-domain detection visualizations. We argue that domain shifts not only cause object representation shifts (leading to false detections), but also increase the similarity between background and object features (resulting in missed detections). These issues highlight the significant challenge of ensuring robust model performance in complex environments across diverse domains.

4. Limitation

Our approach builds upon the CLIP architecture, whose capacity for domain representation is to some extent constrained by the expressive power of the underlying model. Moreover, while OD-LVIS offers a diverse evaluation benchmark, part of its data is augmented through synthetic methods such as style transfer and degradation simulation, which cannot fully capture the full complexity of real-world scenarios. We believe these issues provide fertile ground for future research, including the design of more robust prompt generation mechanisms, the development of stronger vision–language models, and the construction of larger and more realistic multi-domain datasets.

References

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199, 2014. 1

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 2

[3] Jan Flusser, Sajad Farokhi, Cyril Höschl, Tomáš Suk, Barbara Zitova, and Matteo Pedone. Recognition of images degraded by gaussian blur. *IEEE transactions on Image Processing*, 25(2):790–806, 2015. 1

[4] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1

[5] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[6] Stamatios Lefkimmiatis. Universal denoising networks: A novel cnn-based network architecture for image denoising. *arXiv preprint arXiv:1711.07807*, 2017. 1

[7] Mohsen Ebrahimi Moghaddam. A mathematical model to estimate out of focus blur. In *2007 5th International Symposium on Image and Signal Processing and Analysis*, pages 278–281, 2007. 1

[8] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 769–777, 2015. 1

[9] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14824–14834, 2025. 3

[10] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 3

[11] Changfeng Yu, Shiming Chen, Yi Chang, Yibing Song, and Luxin Yan. Both diverse and realism matter: Physical attribute and style alignment for rainy image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12387–12397, 2023. 1

[12] Jing Zhang, Yang Cao, Shuai Fang, Yu Kang, and Chang Wen Chen. Fast haze removal for nighttime image using maximum reflectance prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7418–7426, 2017. 1