Supplementary Material: Towards Open-Vocabulary Video Instance Segmentation

Haochen Wang*
University of Amsterdam
h.wang3@uva.nl

Cilin Yan*
Beihang University
clyan@buaa.edu.cn

Shuai Wang
University of Amsterdam
s.wang3@uva.nl

Xiaolong Jiang, Xu Tang, Yao Hu
Xiaohongshu Inc.
laige@xiaohongshu.com

Weidi Xie†
Shanghai Jiao Tong University
weidi@robots.ox.ac.uk

Efstratios Gavves
University of Amsterdam
egavves@uva.nl

Appendix

A. Dataset Statistic

Statistic Comparison. As shown in Tab. 1, we compare the proposed LV-VIS dataset to existing ones on detailed statistics.

The benchmarks Kitti-MOTS [8], BDD [12], and MOTS-challenge [8] focus on the automatic driving scenario, which contains long videos of street scenes captured from a driving vehicle or a walking pedestrian. The benchmarks DAVIS2017 [6] and Youtube-VOS [10] mainly focus on semi-supervised video object segmentation [6]. The semi-supervised video object segmentation aims to track and segment the objects given the mask of the first frame, which is similar to the single object tracking in VOT [4]. UVO [9] is designed for exhaustively segmenting and tracking anything which humans would consider to be "objects" in videos. Instead of assigning a category label to a specific object, UVO considers all the foreground objects as a single category (in v1 of UVO, the category labels of objects belonging to MS-COCO are annotated, while the rest objects remain unlabeled, as shown in Fig. 2). The category labels are not provided or considered during the evaluation in the above-mentioned datasets.

The Youtube-VIS2019 [11], OVIS [7], and BURST [1] assign a category label to each annotated object. Therefore, category-wise evaluation is enabled. However, the Youtube-VIS2019 and OVIS only contain 40 and 25 categories, which is not wide enough for the open-vocabulary evaluation. The BURST contains a relatively large vocabulary set of 482 categories, but 81% of the object instances in BURST are from MS-COCO categories, making it inappropriate for the evaluation of novel categories, as shown in Fig. 2. Therefore, BURST only considers the 80 common categories in MS-COCO in the category-wise evaluation and mainly follows the evaluation protocols of class-agnostic multiple object tracking.

By contrast, our LV-VIS dataset not only contains a large vocabulary set of 1,196 categories but has a diverse category and object instances distribution, as shown in Fig. 2. Specifically, 94% categories in LV-VIS are disjointed with categories in MS-COCO, while 46% are disjointed with frequent/common categories in LVIS. Moreover, OV-VIS contains 4,828 videos and 544,451 annotated masks for evaluation, which is much larger than most of the validation/test sets in existing datasets. The dataset split detail of LV-VIS is shown in Tab. 2. As shown, we divide LV-VIS into a training set, a validation set, and a test set, where the test set is relatively more complex than the validation set.

Category Partition. In this section, we first illustrate how we select the 1,196 object categories. We first include all 1,203 categories in LVIS. Then we manually select novel categories in ImageNet21K to cover diverse types of categories, such as animals, plants, vehicles, tools, clothing, food, *etc.* In this way, we get 1,612 category candidates. As for categories with multiple meanings, we manually add additional descriptions like LVIS dataset, e.g., date_(fruit), triangle_(musical_instrument). There exist two types of relationships between each of the category pairs: disjoint and non-disjoint. Non-disjoint category pairs could be in partially overlapping, parent-child, or equivalent relationships, which means a single object could have multiple valid category labels. Therefore, we first manually merge object categories with the same semantics, ensuring that there are no mutually equivalent category pairs. In this way, we collect 1,196 well-defined mutually different object categories. Then we construct a parent-child relationship tree among the collected categories. For instance, "race car" is a child of "car". Finally, we revise the annotation based on the defined parent-child relationship tree, ensuring that each object in videos is exhaustively annotated as all corresponding object categories. For instance, a "race car" is also assigned

*Equal contribution.
†Corresponding author.

	Basic			Train				Val/Test			
Dataset	Category	Length(h)	Mask/Frame	Video	Instance	Ann. Frame	Mask	Video	Instance	Ann. Frame	Mask
VOT [4]	-	10.7	1	0	0	0	0	62	62	19,903	19,903
KITTI-MOTS [8]	2	39	5.4	21	748	8,008	38,197	28	961	11,095	61,904
MOTS-Chal. [8]	1	34.4	10	4	228	2,864	26,894	4	328	3,044	32,369
BDD [12]	7	40	11.4	154	17,838	30,745	347,442	32	4,873	6,475	77,389
DAVIS17 [6]	-	2.9	2.6	60	144	4,219	10,238	90	242	6,240	16,841
YT-VOS19 [10]	-	4.5	1.6	3741	6,459	94,400	12,918	1,048	2,115	28,825	4,310
UVO [9]	80*	3	12.3	5,641	76,627	39,174	416,001	5,587	28,271	18,966	177,153
YT-VIS19 [11]	40	4.5	1.7	2,238	3,774	61,845	103,424	645	1,092	17,415	29,431
OVIS [7]	25	3.2	4.7	607	3,579	42,149	206,092	297	1,644	20,492	89,841
BURST [1]	482	28.9	3.1	500	2,645	107,144	318,200	2,414	13,444	88,569	281,957
LV-VIS (Ours)	1196	6.2	4.9	3,083	16,060	70,242	339,533	1,745	9,526	41,253	204,918

Table 1. Detailed Statistic Comparison between our LV-VIS and other video-level datasets. **Category**: The number of the overall category set in each dataset. The - in the Category column means the dataset does not provide the category label or take the category into account during evaluation. **Length**: The total length of videos. **Mask/Frame**: Average annotated masks per frame. Statistics for Val/Test on YT-VIS2019, YT-VOS2019, and DAVIS2017 are estimated from the training set, which may not be exact.

Split	Length(h)	Mask/Frame	Video	Instance	Ann. Frame	Mask
Train	3.9	4.8	3,083	16,060	70,242	339,533
Val	1.1	4.0	838	3,646	19,176	76,916
Test	1.3	5.7	908	5749	22,096	124,834
Total	6.2	4.9	4,828	25,588	111,495	544,451

Table 2. Dataset Split of the LV-VIS.

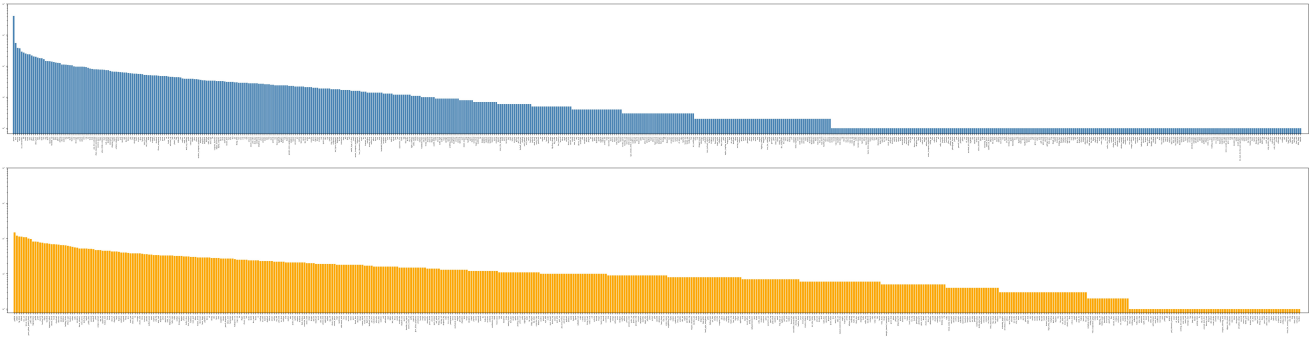

Figure 1. The instance per category on LV-VIS dataset. The blue bars indicate the base categories (frequent/common categories in LVIS) and the orange bars indicate the novel categories (disjointed with base categories).

to the corresponding parent label "car". With the aforementioned pipeline, we address the annotation issues of equivalent categories, parent-child categories, and partially overlapping categories. The category partition and the number of instances per category of the LV-VIS dataset are shown in Tab. 1, where the 1,196 categories are divided into 641 base categories and 555 novel categories. All the base categories in LV-VIS are inherited from frequent/common categories in LVIS [3]. While conducting the evaluation, the categories in Youtube-VIS2019 are divided into 33 base categories and 7 novel categories. The categories in Youtube-VIS2021 are divided into 34 base categories and 6 novel categories. We show the category partitions of Youtube-VIS2019 and Youtube-VIS2021 [11] in Tab. 3.

Annotation Details. We develop a video segmentation annotation platform based on Labelme, which is released here https://github.com/haochenheheda/segment-anything-annotator. We first manually annotate all object masks in the first frame by polygons and propagate the object masks to the second frame with STCN [2]. After that, we correct the propagated masks, add masks for newly appeared objects, and then repeat the propagation to the next frame. We manually recognize and assign category names to each annotated mask sequence. Finally, we include cross-revision to ensure the annotation quality.

B. Visualizations of Annotated Frames

Examples of annotated videos in LV-VIS are shown in Fig. 3.

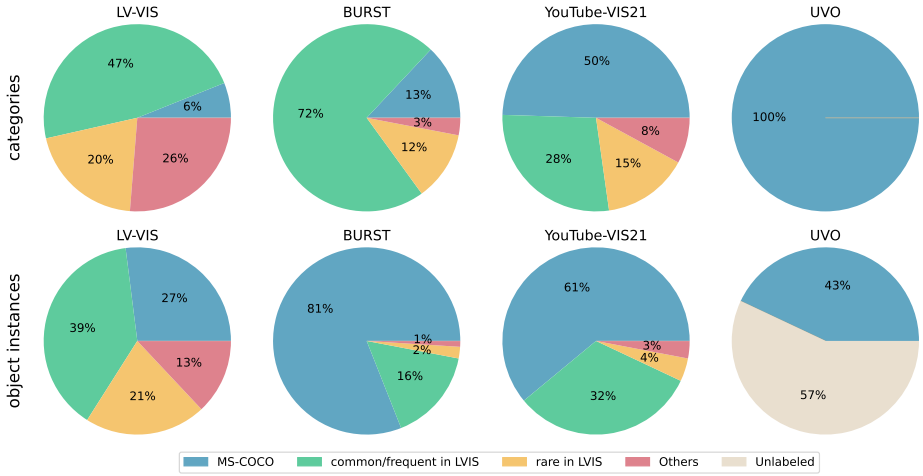

Figure 2. Category and object instance distributions. To avoid overlap, categories from MS-COCO are excluded in common/frequent LVIS to draw the figures. A significant proportion of categories and objects in the LV-VIS are distinct from the commonly used datasets.

Figure 3. Sample videos in LV-VIS.

C. Visualization and Failure Case Analysis

We demonstrate the result of OV2Seg on our proposed LV-VIS, Youtube-VIS [11] and OVIS [7] in Fig. 5. OV2Seg shows a strong generalization ability on those video instance segmentation datasets, even in some hard cases, to be specific, (b) large perspective change, (c) blurry video, (f) long video, (a, g) occlusion, and (g) a large number of objects.

The failure cases are demonstrated in Fig 5 (h)-(i). The major failure case is category confliction, which means the classification of objects from novel categories is usually dominated by their visually similar base categories. To be specific, as demonstrated in Fig 5 (i), The ”wolf” in the figure is recognized as a ”dog” because of the apparent similarity. As the objects of ”dog” are shown in the training set while the objects from ”wolf” are not, the model learns better alignments between the object embedding and the word embedding of ”dog,” which makes the model tend to recognize an object as ”dog” instead of a ”wolf.” We consider the category confliction as a fundamental challenge for all the open-vocabulary tasks, which could be improved by including a large vocabulary set during training or some training protocol to transfer the information from the image domain, such as knowledge distillation or self-training. Another failure case is the miss segmentation of some common categories, such as the ”person” in Fig. 5 (b), (h), (i). This is because the LVIS is not a densely annotated dataset. Only a part of the objects are annotated, especially for the most common objects such as ”person.” Specifically, LVIS only annotated 13,439 ”persons” out of 262,465 (annotated in MS-COCO [5]) in total. Therefore most of the ”person” objects in the training set are regarded as background, which leads to a low recall of the person category. This could be relieved by involving some semi-supervised training methods or combining the LVIS with the MS-COCO [5] dataset to fill the miss annotations of the common categories in LVIS. We hope to inspire future works by giving analyses of the failure cases.

Datasets	Type	Categories						
Youtube-VIS2019	Base	airplane	bear	boat	cat	cow	deer	dog
		duck	eagle	elephant	fish	frog	giant panda	giraffe
		horse	lizard	monkey	motorbike	mouse	owl	parrot
		person	rabbit	shark	skateboard	snowboard	surfboard	tennis racket
		tiger	train	truck	turtle	zebra		
	Novel	earless seal	fox	leopard	snake	ape	hand	sedan
Youtube-VIS2021	Base	airplane	bear	boat	car	cat	cow	deer
		dog	duck	eagle	elephant	fish	frog	giant panda
		giraffe	house	lizard	monkey	motorbike	mouse	owl
		parrot	person	rabbit	shark	skateboard	snowboard	surfboard
		tennis	tiger	train	truck	turtle	zebra	
	Novel	earless seal	fox	leopard	snake	flying dsic	whale	

Table 3. Base and novel categories in Youtube-VIS2019 and Youtube-VIS2021 datasets.

Figure 4. A screenshot of annotated frames in LV-VIS. The full videos and annotations will be released upon publication.

References

[1] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 2023. 1, 2

[2] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2

[3] 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. 2

[4] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 0–0, 2018. 1, 2

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 3

[6] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 2

[7] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 2022. 1, 2, 3

[8] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7942–7951, 2019. 1, 2

[9] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10776–10785, 2021. 1, 2

[10] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1, 2

[11] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 1, 2, 3

[12] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020. 1, 2

Figure 5. Predictions of OV2Seg on Video Instance Segmentation datasets. Figures (a), (b), and (i) are from the LV-VIS dataset. Figures (c), (d), (e), and (h) are from Youtube-VIS2019/2021 datasets. Figures (f) and (g) are from the OVIS dataset. Figures (h) and (i) are failure cases.