OVT-B: A New Large-Scale Benchmark for Open-Vocabulary Multi-Object Tracking

Haiji Liang[1], Ruize Han[2,3]*

[1] School of Software Technology, Zhejiang University
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3] Department of Computer Science, City University of Hong Kong

GitHub

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Open-vocabulary object perception has become an important topic in AI, which aims to identify objects with novel classes that have not been seen during training. Under this setting, open-vocabulary object detection (OVD) in a single image has been studied in many literature. However, the **open-vocabulary object tracking (OVT)** from a video is less studied, and a reason is the shortage of benchmarks. In this work, we build OVT-B, a large-scale **O**pen-**V**ocabulary multi-object **T**racking **B**enchmark. OVT-B contains 1,048 categories of objects and 1,973 videos with 637,608 bounding box annotations, which is much larger than the sole open-vocabulary tracking dataset, *i.e.*, OV-TAO-val dataset (200+ categories, 900+ videos). **The proposed OVT-B can be used as a new benchmark to pave the way for the research of OVT**. We also develop a simple yet effective baseline method OVTrack+, which integrates the motion features for OVT task.

OVT Benchmark

Comparison with MOT datasets

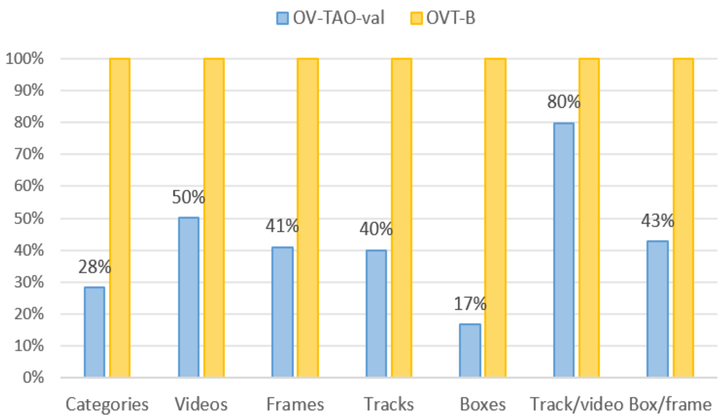
OVT-B dataset comprises 1,048 categories (534 base and 514 novel). In comparison to existing MOT datasets, such as MOT17, KITTI, and UAVDT-MOT, OVT-B offers a significantly larger variety of categories, surpassing even the TAO dataset with 833 categories. OVT-B offers a larger number of annotated frames, trajectories, bounding boxes, and videos, making it a dataset of significantly greater scale.

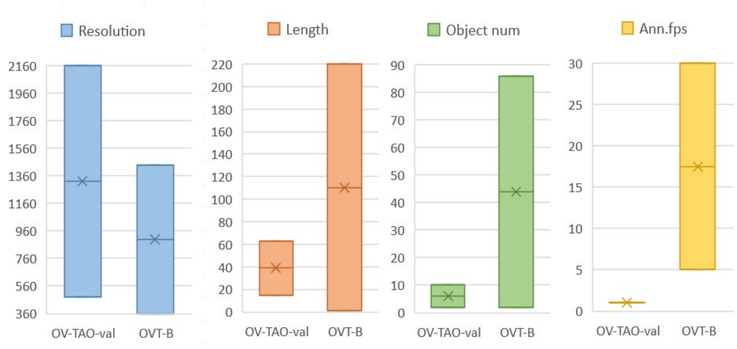
Word cloud of OVT-B categories

Statistics of MOT datasets and OVMOT datasets

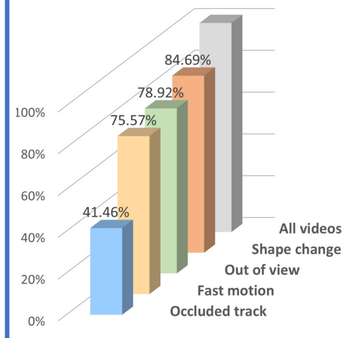
Datasets	#Cls.	#Vid.	#Track	#Box	#Frm.	Res.	Dur.	#Obj.	Ann.
MOT17	1	42	3993	901K	33K	480-1080	17-85	1-63	30
MOT20	1	8	3833	2102K	13K	880-1080	17-133	1-94	30
KITTI	5	50	2600	80K	15K	512	20-90	0-30	10
DanceTrack	1	100	990	877K	105K	720-1080	20-108	1-22	20
UAVDT	3	100	2700	841K	80K	540-1080	3-99	1-122	6
TAO	833	2907	17287	333K	2674K	480-2160	1-279	1-10	1
GMOT-40	10	40	2026	256K	9K	480-1080	3-24.2	10-128	24-30
OV-TAO-val	330	988	5473	113K	36K	480-2160	15-63	1-11	1
OV-TAO-test	357	1419	7946	166K	52K	480-2160	10-59	1-11	1
OVT-B (Ours)	1048	1973	13686	673K	88K	360-1440	1-220	2-86	5-30

Comparison of OV-TAO-val and OVT-B





Comparison with OV-TAO-val: As a rigorously developed benchmark, OVT-B significantly surpasses the OV-TAO-val dataset, offering dense targets, comprehensive annotations, and a diverse range of videos.



Ratio of videos with attributes

Dataset Attributes

OVT-B presents various challenges for tracking, including objects moving out of view, rapid motion, shape changes, and varying levels of occlusion. Besides, OVT-B contains a proportion of large objects, objects with complex shapes, and those with short trajectories. These attributes highlight the diversity of targets and trajectories present in OVT-B.

Screenshots of annotations of OVT-B

OVTrack+: A New Baseline

To tackle the challenge of open-vocabulary multi-object tracking, we propose OVTrack+ integrating the motion model for the object association task in OVT, thanks to its category-agnostic nature.

Method	All				Base				Novel			
	TETA	LocA	AssA	ClsA	TETA	LocA	AssA	ClsA	TETA	LocA	AssA	ClsA
ByteTrack [58]	20.1	36.1	12.4	11.9	20.6	35.6	12.7	13.4	19.6	36.6	12.0	**10.3**
OC-SORT [59]	16.0	31.2	4.3	**12.3**	16.5	31.0	4.4	**14.3**	15.4	31.4	4.3	**10.3**
StrongSORT [14]	24.8	31.6	30.7	12.2	25.7	31.4	31.6	14.2	23.9	31.8	29.7	**10.3**
OVTrack [5]	46.1	60.8	66.1	11.5	46.8	60.5	66.7	13.4	45.5	61.1	65.5	9.6
OVTrack+	**47.0**	**62.0**	**67.7**	11.3	**47.6**	**61.6**	**68.2**	13.2	**46.4**	**62.5**	**67.3**	9.4

Open-vocabulary MOT comparison results on OVT-B

Method	All				Base				Novel			
	TETA	LocA	AssA	ClsA	TETA	LocA	AssA	ClsA	TETA	LocA	AssA	ClsA
ByteTrack [58]	20.1	36.9	6.0	**17.6**	20.9	37.0	5.9	**19.7**	14.7	36.0	6.1	1.8
OC-SORT [59]	24.3	52.1	6.0	14.8	25.1	52.7	6.1	16.5	18.5	48.1	5.4	**2.1**
StrongSORT [14]	23.4	41.6	13.5	15.2	24.4	42.3	13.7	17.0	16.6	36.4	11.6	1.7
OVTrack [5]	36.1	53.8	37.3	17.3	37.1	54.2	37.8	19.4	28.8	51.2	33.7	1.5
OVTrack+	**38.4**	**57.5**	**40.8**	16.9	**39.2**	**57.5**	**41.0**	18.9	**32.5**	**57.0**	**38.7**	1.8

Open-vocabulary MOT comparison results on OV-TAO-val