

VastTrack: Vast Category Visual Object Tracking

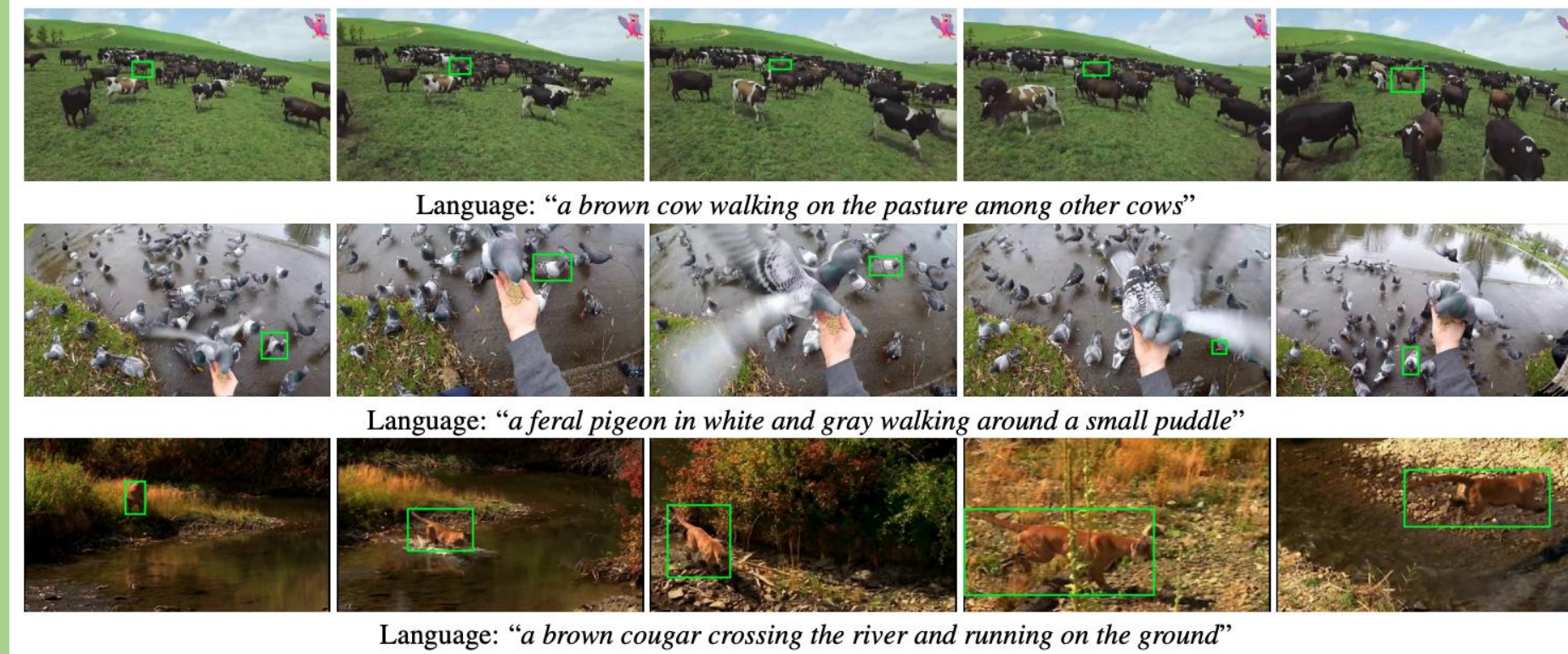
Liang Peng^{1*} Junyuan Gao^{1*} Xinran Liu^{1*} Weihong Li^{1*} Shaohua Dong^{2*} Zhipeng Zhang³ Heng Fan^{2†} Libo Zhang^{1†}
(*equal contributions; †equal advising and co-last author)

¹Institute of Software Chinese Academy of Sciences

²University of North Texas

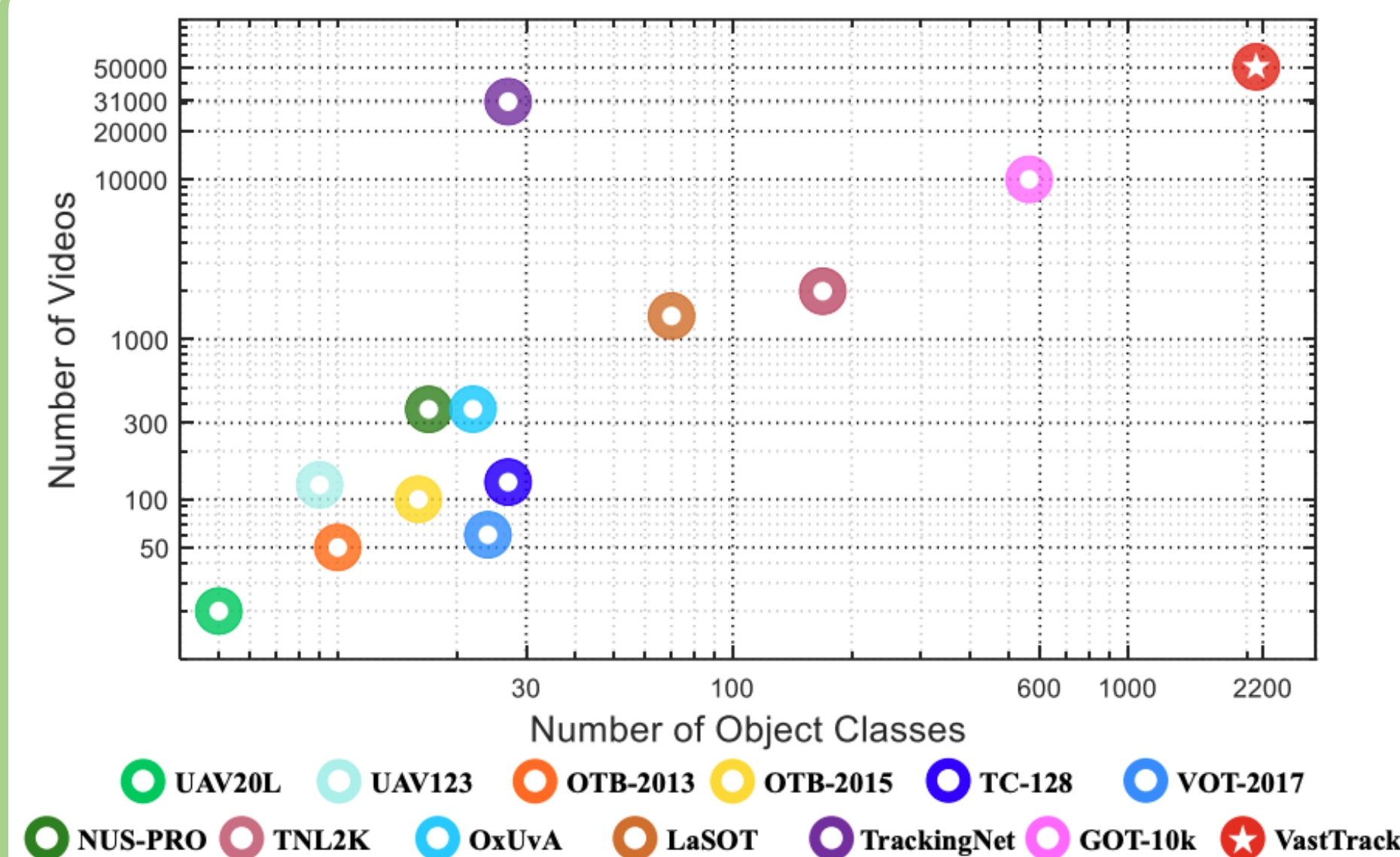
³KargoBot

Introduction



- We propose VastTrack, the largest tracking benchmark regarding the number of videos and object categories.
- We offer high-quality comprehensive labeling, including bounding box annotation and natural language description for each video.
- We evaluate 25 state-of-the-art visual trackers on VastTrack, providing extensive baselines for future comparisons.

Motivation

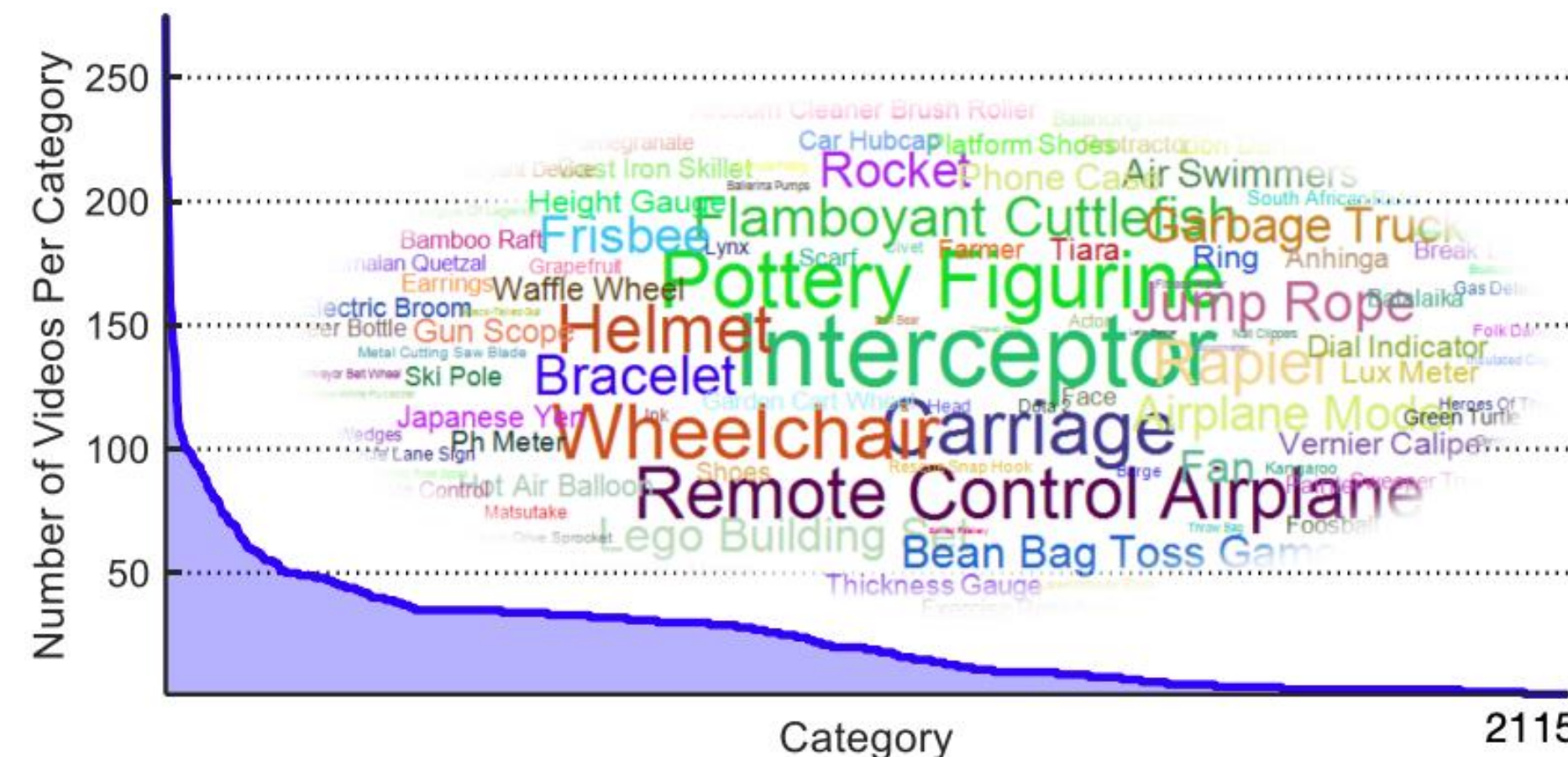
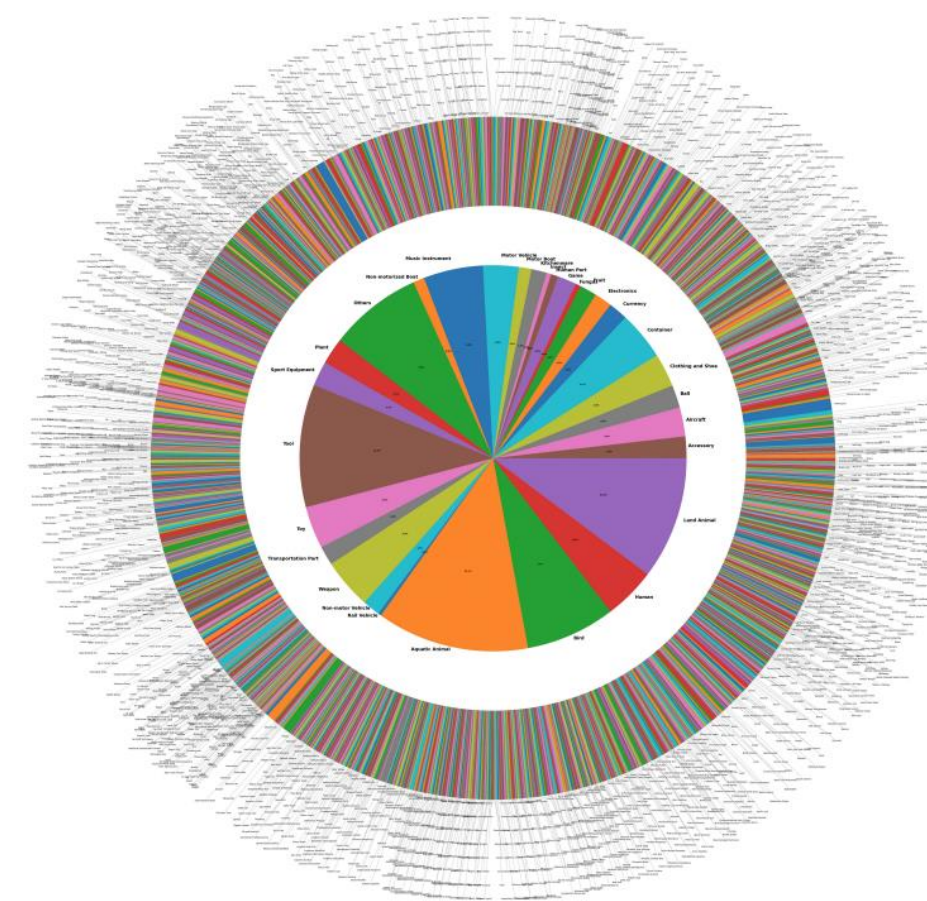


- Limitations of existing tracking benchmarks
 - **Restricted number of categories**
 - **Limited scale for training generic trackers**
 - **Small number of natural language descriptions**

Proposed VastTrack

Benchmark	Year	Classes	Videos	Mean Frames	Total Frames	Total Duration	Absent label	Num. of Att.	Lang. Anno.	Frame Rate	Dataset Focus	Dataset Goal
OTB-2013 [53]	2013	10	50	578	29K	16.4 min	✗	11	✗	30 fps	ST	Eva.
OTB-2015 [54]	2015	16	100	590	59K	32.8 min	✗	11	✗	30 fps	ST	Eva.
TC-128 [36]	2015	27	128	429	55K	30.7 min	✗	12	✗	30 fps	ST	Eva.
NUS-PRO [32]	2016	17	365	371	135K	75.2 min	✗	12	✗	30 fps	ST	Eva.
UAV123 [42]	2016	9	123	915	113K	62.5 min	✗	12	✗	30 fps	ST	Eva.
UAV20L [42]	2016	5	20	2,934	59K	32.6 min	✗	12	✗	30 fps	LT	Eva.
NfS [21]	2017	17	100	3,830	383K	26.6 min	✗	9	✗	240 fps	ST	Eva.
VOT-2017 [29]	2017	24	60	356	21K	11.9 min	✗	24	✗	30 fps	ST	Eva.
OxUvA [46]	2018	22	366	4,235	1.55M	14.4 hours	✗	6	✗	30 fps	LT	Eva.
TrackingNet [43]	2018	27	30,643	471	14.43M ^b	140.0 hours	✗	15	✗	30 fps	ST	Tra./Eva.
LaSOT [16]	2019	70	1,400	2,053	3.52M	32.5 hours	✓	14	✓	30 fps	LT	Tra./Eva.
TNL2K [50]	2021	169 [†]	2,000	622	1.24M	11.5 hours	✓	17	✓	30 fps	ST	Tra./Eva.
GOT-10k [27]	2021	563	9,935	149	1.45M	40.0 hours	✓	6	✗	10 fps	ST	Tra./Eva.
VastTrack	2024	2,115	50,610	83	4.20M	194.4 hours	✓	10	✓	6 fps	ST	Tra./Eva.

- VastTrack



- ❖ **Vast Object Category:** VastTrack covers targets from 2,115 categories, significantly surpassing classes of existing benchmarks (e.g., GOT-10k with 563 classes and LaSOT with 70 categories). It is the richest dataset to date.
- ❖ **Larger scale:** VastTrack offers 50,610 videos with 4.2 million frames, which makes it the largest and the most diverse tracking dataset in terms of the numbers of videos and targets compared to existing datasets.
- ❖ **Rich and Precise Annotations:** VastTrack offers both standard bounding box annotations and rich linguistic specifications for videos, and thus enables exploration of both vision-only and vision-language tracking.

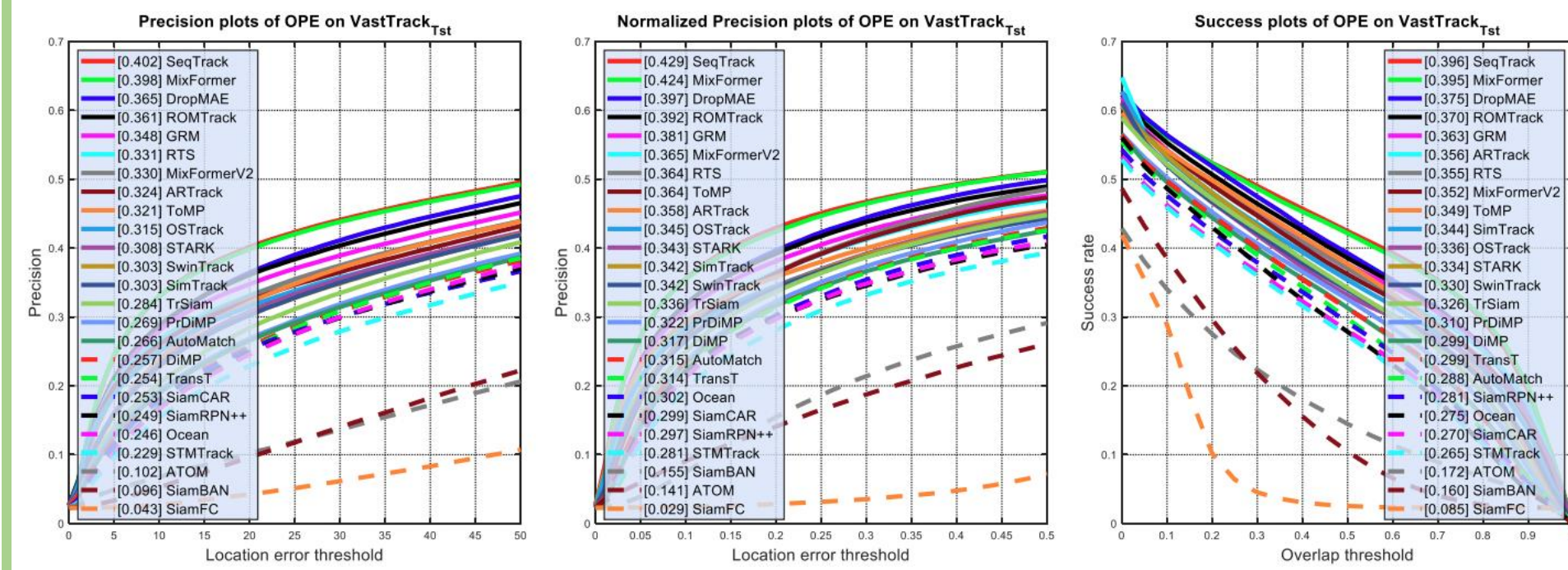
- Dataset Split and Evaluation Protocol

	Classes	Videos	Mean frames	Total frames
VastTrack _{Tst}	702	3,500	106.3	372K
VastTrack _{Tra}	1,974	47,110	81.2	3.82M

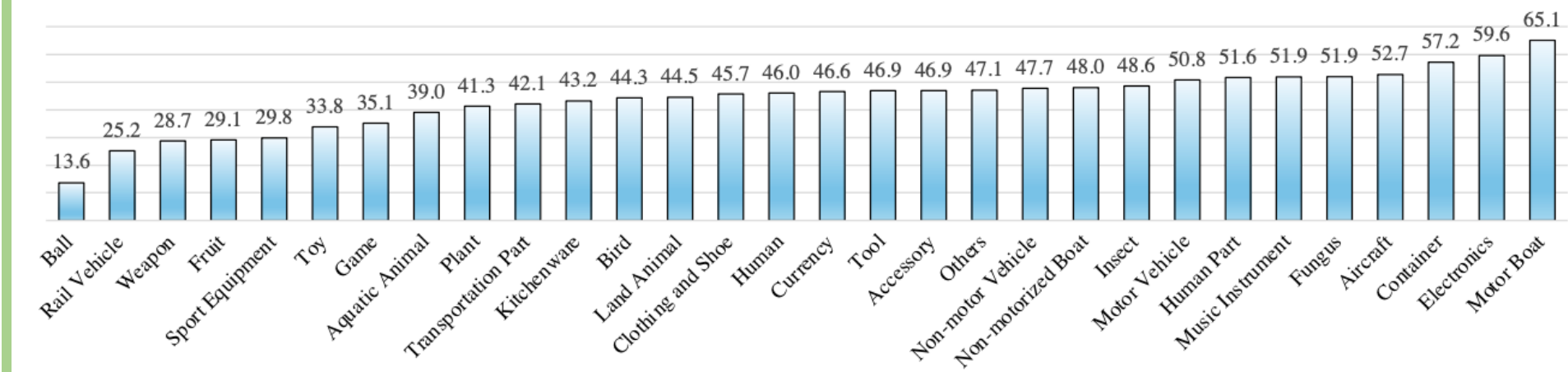
- ❖ **Evaluation Protocol:** We utilize a hybrid protocol wherein part of object classes (not videos) in test set have overlap with training set, while the rest classes remains unseen.

Experiments

- Evaluation of 25 state-of-the-art-trackers using precision, normalized precision and success scores.



- Comparison on meta classes using success score



- Comparison of VastTrack with other datasets

	Success Score			
	TrackingNet [43]	LaSOT [16]	TNL2K [50]	VastTrack (Ours)
SeqTrack [6]	0.855	0.725	0.578	0.396
MixFormer [9]	0.854	0.724	0.533	0.395
DropMAE [52]	0.841	0.718	0.569	0.375
ROMTrack [4]	0.841	0.714	0.604	0.370
GRM [22]	0.840	0.699	0.611	0.363
ARTrack [51]	0.843	0.708	0.575	0.356
RTS [44]	0.816	0.697	0.599	0.355
MixFormerV2 [10]	0.834	0.706	0.506	0.352
ToMP [40]	0.815	0.685	0.584	0.349
SimTrack [5]	0.834	0.705	0.556	0.344
OSTrack [57]	0.839	0.711	0.559	0.336
STARK [56]	0.820	0.671	0.525	0.334
SwinTrack [38]	0.811	0.672	0.559	0.330
TrSiam [49]	0.781	0.624	0.523	0.326
PrDiMP [12]	0.758	0.598	0.470	0.310

Conclusion

- We propose a new large-scale dataset VastTrack for vast category tracking
- Our experiments show that there is still a long way for generic tracking

