Beyond SOT: Tracking Multiple Generic Objects at Once - Supplementary

Christoph Mayer[1*] Martin Danelljan[1] Ming-Hsuan Yang[2] Vittorio Ferrari[2]
Luc Van Gool[1] Alina Kuznetsova[2]
[1]ETH Zürich [2]Google Research
{christoph.mayer, martin.danelljan, vangool}@vision.ee.ethz.ch {minghsuan, vittoferrari, akuznetsa}@google.com

In this supplementary material, we first give an overview of the different task definitions and the corresponding abbreviations used in the paper and supplementary material. Next, we describe the details of the model architecture and training in Sec. 2. Then, we provide more insights into the experiments presented in the main paper and provide additional results on less popular tracking datasets in Sec. 3. Then, we show visual results between the baseline and our tracker on multiple sequences of the proposed datasets including failure cases 4. Next, we discuss the limitations of the proposed tracker in Sec. 5. Finally, we provide additional insights about our dataset and compare it to datasets of related tasks in Sec. 6.

1. Glossary

In this Section we will briefly summarize the different task definitions behind the individual abbreviations:

GOT. Generic Object Tracking refers to the task of tracking potentially multiple user-defined target objects of arbitrary classes specified by a user-specified bounding box in the initial video frame.

SOT. Singe Object Tracking is the same task as GOT but focuses on the setting where only a single generic object needs to be tracked.

Multi-Object GOT. The same as GOT but emphasizes that multiple-objects need to be tracked. We use multi-object GOT because GOT is in other research works sometimes used interchangeably with SOT.

MOT. Multi Object Tracking is completely different from the tasks listed above because it requires a class category list to detect and track all objects corresponding to the defined class categories.

GMOT. Generic Multi Object Tracking is the same as MOT but instead of using a class category list to define the target objects, a single user-specified box shows an example object of the target class category. Thus, all objects that belong to the same class as the user-specified example need to be detected and tracked.

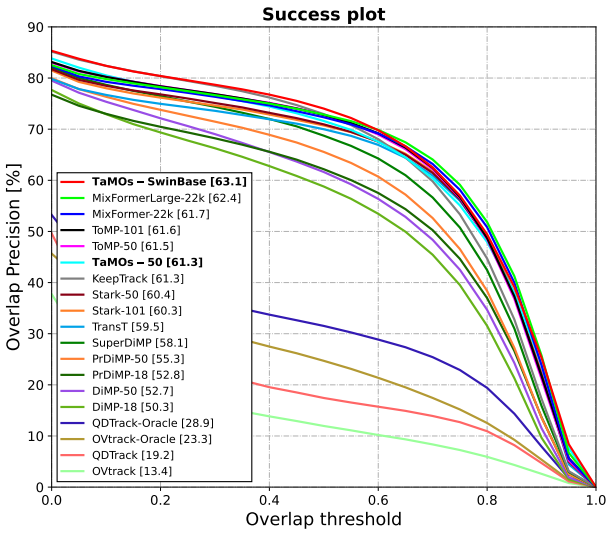
*Work done while interning at Google Research.

2. Model Architecture and Training Details

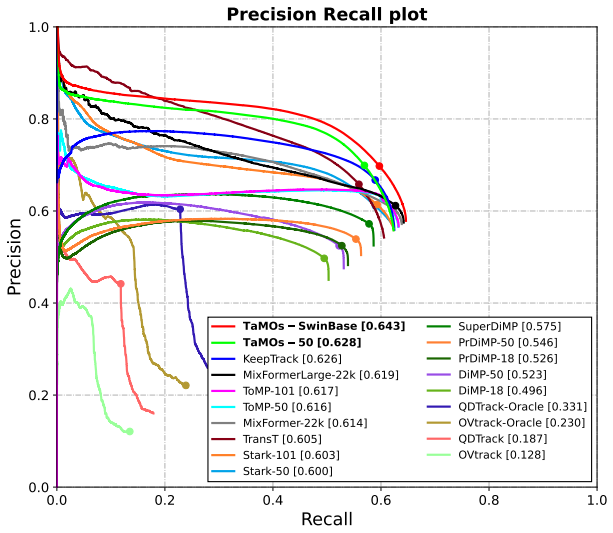
Architecture. We extract backbone features either from the ResNet-50 or the SwinBase backbone. For both backbones we extract the features corresponding to the blocks with stride 8 and 16. We only use the features with stride 16 for object encoding and feed these features into the model predictor. For both backbones we use a linear layer to decrease the number of channels from 1024 to 256 or 512 to 256 respectively. Thus we use 256 dimensional object embeddings e_i and a Multi-Layer Perceptron (MLP) to project the LTRB bounding box encoding map from 4 to 256 channels. Since the model predictor produces 256 dimensional convolutional filters we require the same number of channels for the Feature Pyramidal Network (FPN) output features. In particular we use a two layer FPN that uses as input the enhanced Transformer encoder output features corresponding to the test frame as well as the aforementioned high resolution backbone test features. The high resolution input features have either 512 or 256 channels for the Resnet-50 or the SwinBase backbone respectively. Thus, we adapt the FPN accordingly depending on the used backbone.

Training Details. Since our tracker operates on full frames, we retain the full training and testing frames. The frames are re-scaled and padded to a resolution of 384×576. As we use the feature maps with stride 16 for both the ResNet-50 [14] and SwinBase [18] backbones, this results in an extracted feature and score map resolution of 24×36. For ResNet-50 we use pretrained weights on ImageNet-1k and for SwinBase on ImageNet-22k. We use a fixed size Gaussian when producing the score map encoding for each object where $\sigma = 0.25$. Furthermore, we use gradient norm clipping with the parameter 0.1 in order to stabilize training. In addition, we employ data augmentation techniques during training such as random scaling and cropping in addition to color jittering and randomly flipping the frame. The regression loss is given by

$$L_{\text{bbreg}} = \sum_{i=0}^{n} L_{\text{GIoU}} \left(\hat{b}_i^{\text{ltrb}}, \hat{b}_i^{\text{ltrb}} \right), \tag{1}$$



(a) Success Plot



(b) Precision-Recall

Figure 1. Success plot, showing OP_T, on LaGOT (AUC is reported in the legend). Tracking Precision-Recall curve on LaGOT – VOTLT is reported in the legend (the highest F1-score).

where L_{GIoU} denotes the generalized IoU-Loss [26]. The overall training loss is then defined as

$$L_{\text{tot}} = \lambda_{\text{cls}} L_{\text{cls}}(\hat{y}, y) + \lambda_{\text{bbreg}} \cdot L_{\text{bbreg}}(\hat{b}^{\text{ltrb}}, b^{\text{ltrb}}) \quad (2)$$

where $\lambda_{\text{cls}} = 100$ and $\lambda_{\text{bbreg}} = 1$ are scalars weighting the contribution of each loss component. We use ADAMW [19] with a learning rate of 0.0001 that we decay after 150 and 250 epochs by a factor of 0.2 and train all models on four A100 GPUs with a batch size of 4×12 or 4×6.

Inference. During inference we adopt the simple memory updating approach described in [20]. In particular, updating the memory refers to adding a second dynamic training frame using predicted box annotations. We replace the second training frame (update the memory) if the maximal value in each score map is above the threshold of $\tau = 0.85$.

For accurate bounding box prediction and localization we employed an FPN. In contrast to training, where we applied the target models directly on the Transformer encoder features and also on the low- and high-resolution FPN feature maps, we only use the high-resolution score and bounding box prediction maps during inference. We empirically observed better training performance when applying the losses on each instead of only on the high resolution outputs. However, during inference we are only interested in the high resolution predictions.

3. Experiments

We provide more detailed results to complement the comparison shown in the main paper. In addition we provide result for the LaSOTExt [10] dataset in order to assess the performance of our tracker on sequences containing small objects. Similarly, we analyze the capability of our tracker to handle adverse tracking conditions on AVisT [24]. Furthermore, to provide results on another multiple object dataset we run the tracker on ImageNetVID [27].

Table 1. Comparison of the combination of GOT and MOT methods. GOT return the detections and the MOT methods are used for object association over time on LaGOT.

GOT	MOT	F1-Score	Success	HOTA	MOTA	IDF1
TaMOs-SwinBase	—	0.643	63.1	62.1	58.2	74.7
	SORT [2]	0.438	35.7	45.9	52.2	43.3
	ByteTrack [33]	0.459	37.7	50.4	57.1	53.9
MixFormerLarge-22k	—	0.619	62.4	61.5	52.3	74.3
	SORT [2]	0.418	34.0	45.6	43.9	44.9
	ByteTrack [33]	0.450	36.4	47.5	44.8	49.6

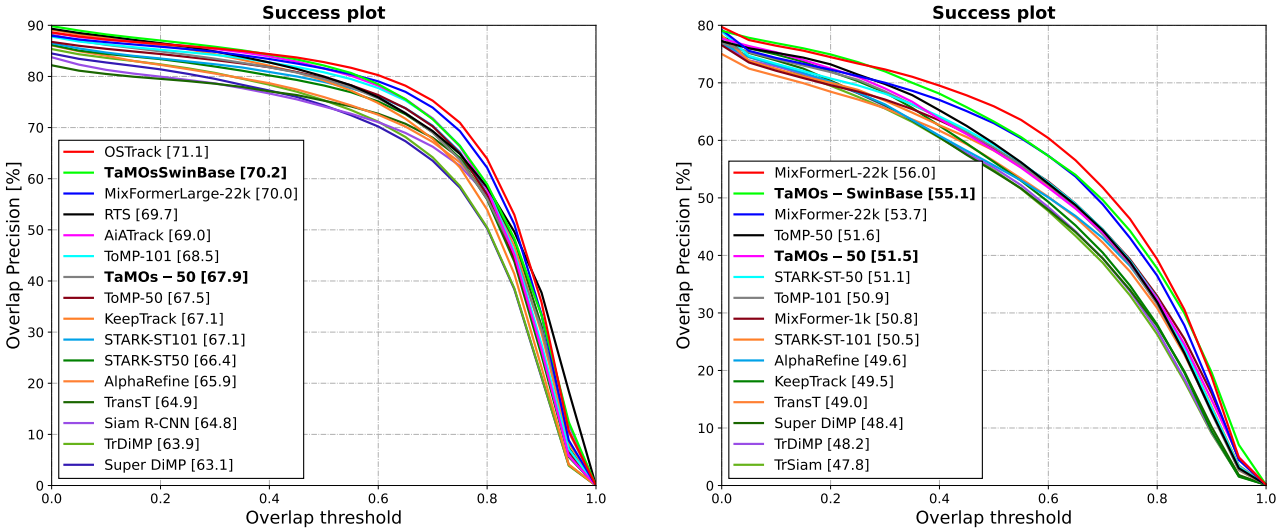
3.1. LaGOT

To complement the results shown in the main paper, we report in Fig. 1 and Tab. 2 results for additional trackers and different variants, such as using a different backbone or different hyper-parameters. In Tab. 2 we report additional Multiple Object Tracking (MOT) sub-metrics and statistics on LaGOT. In general we conclude, that using larger backbones especially if they are pretrained on ImageNet-22k leads to the best results. Furthermore, we observe that the MOT methods QDTrack and OVTrack (evaluated with default parameters provided in the OVTrack GitHub repository[1]) are not competitive with Generic Object Tracking

[1] https://github.com/SysCV/ovtrack

Table 2. Comparison of different trackers using MOT metrics on LaGOT.

		HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	LocA	OWTA	MOTA	IDSW	IDF1
GOT	**TaMOs-SwinBase**	62.1	57.3	68.4	69.9	69.9	75.9	75.9	84.2	68.9	58.2	6734	74.7
	TaMOs-50	60.0	54.6	66.9	67.7	67.7	74.5	74.5	84.0	67.1	52.9	7901	72.0
SOT	MixformerLarge-22k	61.5	53.8	70.9	67.4	67.4	77.8	77.8	84.8	69.0	52.3	3150	74.3
	Mixformer-22k	61.2	54.0	70.0	67.4	67.4	77.0	77.0	84.5	68.6	53.2	3339	74.4
	ToMP-101	60.1	53.0	68.8	66.4	66.4	76.2	76.2	83.9	67.5	51.9	2638	73.8
	ToMP-50	60.0	53.0	68.6	66.4	66.4	76.0	76.1	83.8	67.4	52.3	2378	74.0
	STARK-ST-101	59.4	51.8	68.8	65.6	65.6	75.9	75.9	84.2	67.1	49.0	3568	72.5
	STARK-ST-50	59.4	51.9	68.5	65.6	65.6	75.6	75.6	83.9	66.9	49.5	4277	72.6
	TransT	57.8	50.2	67.1	64.3	64.3	74.5	74.6	84.3	65.6	46.6	2323	70.7
	KeepTrack	59.1	52.3	67.3	65.4	65.4	74.7	74.7	82.3	66.2	51.3	2299	73.8
	SuperDiMP	56.1	48.3	65.8	62.1	62.1	73.5	73.5	82.2	63.8	43.2	1966	69.7
	PrDiMP-50	53.0	45.6	62.1	59.6	59.6	70.3	70.4	81.3	60.7	38.4	2380	66.6
	PrDiMP-18	51.4	42.8	62.2	57.2	57.2	70.2	70.3	81.3	59.5	31.9	1981	63.4
	DiMP-50	50.8	42.1	62.0	56.2	56.2	69.7	69.7	80.2	58.9	29.4	1680	62.1
	DiMP-18	48.1	39.3	59.6	53.5	53.5	67.6	67.6	79.5	56.3	23.2	1757	59.0
MOT	QDTrack	22.2	17.3	29.0	46.2	21.0	30.3	80.0	81.8	36.3	-115.8	18521	16.3
	OVTrack	24.4	20.3	29.9	22.7	59.7	31.2	78.2	82.0	25.9	13.9	4951	23.5



(a) LaSOT

(b) AViST

Figure 2. Success plot, showing OP_T, on LaSOT [11] and AVisT [24] (AUC is reported in the legend).

(GOT) methods. In particular, we observe that QDTrack and OVTrack achieve very low OWTA scores that depend on the Detection Recall (DetRe) and the Association Accuracy (AssA) scores. OVTrack scores the lowest DetRe despite being an open-vocabulary detector. While this is an expected limitation, we further observe that QDTrack achieves by far the lowest AssA caused by the poor Association Recall (AssRe) of 30.3 compared to DiMP-18 that achieves 67.6.

In addition to the Single Object Tracking (SOT) and MOT baselines presented in the main paper, we also evaluate an open-world tracker [17]. Such a tracker aims at tracking all objects in the scene and should therefore also be able to track the generic objects contained in LaGOT. In particular, we follow Liu *et al*. [17] and generate object proposals for each video frame using their provided open-world detector. Then, we run SORT [2] on top of the generated proposals using the default parameters. This leads to an OWTA score of 12.58, AssA of 3.57 and DetRe of 46.72. We conclude that the complex videos with long tracks of the proposed benchmark are for now too challenging for existing open-world trackers.

Finally, we add another experiment where we use our tracker TaMOs or the SOT tracker Mixformer as one-shot object detectors and feed their detections and scores to a MOT tracker that focuses on building the final tracklets. In

Table 3. Comparison to the state of the art on LaSOTExt [10].

Method	Venue	LaSOTExt [10] Prec	N-Prec	Succ
TaMOs-SwinBase		**58.0**	57.8	**49.2**
TaMOs-Resnet-50		54.1	55.0	46.7
AiATrack [13]	ECCV'22	54.7	58.8	49.0
OSTrack [31]	ECCV'22	**57.6**	**61.3**	**50.5**
ToMP-101 [20]	CVPR'22	52.6	58.1	45.9
ToMP-50 [20]	CVPR'22	51.9	57.6	45.4
GTELT [34]	CVPR'22	52.4	54.2	45.0
KeepTrack [21]	ICCV'21	54.7	**61.7**	48.2
SuperDiMP [6]	CVPR'20	49.0	56.3	43.7
LTMU [4]	CVPR'20	45.4	53.6	41.4
DiMP [3]	ICCV'19	43.2	49.6	39.2
ATOM [5]	CVPR'19	41.2	49.6	37.6

Table 4. Analysis of the FPN and the zooming mechanism on LaSOTExt [10] and UAV123 [22].

Backbone	FPN	Zoom	LaSOTExt AUC	UAV123 AUC
Resnet-50	✗	✗	41.3	56.2
Resnet-50	✓	✗	43.1	58.2
Resnet-50	✓	✓	**46.7**	**64.2**
SwinBase	✗	✗	43.9	56.5
SwinBase	✓	✗	44.6	57.3
SwinBase	✓	✓	**49.2**	**66.2**

particular we use the popular SORT [2] tracker and the recent state-of-the-art tracker ByteTrack [33]. For TaMOs and Mixformer, using their predicted bounding boxes and object ids leads to far better results than using an MOT method on top for post-processing. This behaviour holds when measuring the performance of the resulting trackers with GOT as well as with MOT metrics, see Tab. 1. While there is potential to increase the robustness of GOT trackers in case of multiple objects, directly applying MOT trackers is not a good solution. Instead dedicated association algorithms for multi-object GOT are needed. We conclude, that TaMOs and the proposed SOT trackers run in parallel, are solid baselines for LaGOT.

3.2. LaSOT

In addition to the result table, shown in the main paper, we show in Fig. 2a the success plot for LaSOT [11]. We observe that our tracker is the most robust ($T < 0.3$). Furthermore, the plot shows that both MixFormerLarge-22k and OSTrack can regress more accurate bounding boxes ($0.5 < T < 0.9$). However, unlike these specialized single-target object trackers, our approach is capable of jointly tracking multiple targets.

3.3. LaSOTExt

Since our tracker always operates on the full frame without the help of a local search region, tracking small objects is challenging. Thus, we integrated an FPN in our tracker to improve the tracking accuracy. To analyze our tracker on small objects we run it on LaSOTExt [10] and UAV123 [22]. Tab. 4 shows that including an FPN improves the tracking results on both datasets but is more effective when using a Resnet-50 as backbone.

To track small objects a high feature map resolution is desirable. To better cope with extremely small objects, found in some SOT benchmarks, we add a simple zooming mechanism. In particular, when the target is smaller that 30×30 pixels, we crop a region of the image that ensures this minimal target size when up-scaled to the input-resolution of 384×576. Tab. 4 clearly shows that using such a zooming mechanism improves the results on LaSOTExt and UAV123 considerably, due to the presence of extremely small objects in these datasets.

Tab. 3 shows that our tracker with FPN and zooming achieves competitive results on LaSOTExt. In particular it achieves the highest precision and the second highest success AUC only being outperformed by OSTrack [31].

3.4. AVisT

In order to validate our tracker in adverse visibility scenarios we run it on AVisT [24]. Fig 2b shows that our tracker achieves excellent results with a success AUC of 55.1. This result shows that our tracker is able to track generic single objects even in visually challenging scenarios. The best tracker MixFormerLarge-22k is able to regress more accurate bounding boxes ($0.3 < T < 0.9$), as it relies on small search area selection to ensure high-resolution features. In contrast, our approach is capable of jointly tracking multiple objects.

3.5. ImageNetVID

In order to validate the proposed multiple object GOT tracker not only on LaGOT but also on another multiple object dataset, we modify ImageNetVID [27]. Since ImageNetVID is a video object detection datasets instead of a GOT dataset we perform the following adaptations. First, we remove all tracks that are not present in the first frame. Then, we use the remaining tracks to produce the bounding box annotations of the first frame. For simplicity we remove the 11 sequence where no track is visible in the first frame. This results in 544 sequences with 938 tracks and 1.7 tracks on average per video. Fig. 4 shows the success plot on the resulting multiple object GOT dataset. We observe that all trackers achieve relatively high AUC mostly differing in bounding box accuracy. Both versions of our tracker outperform the baselines ToMP-50 and ToMP-101 [20]. In particular, we notice the superior bounding box accuracy of

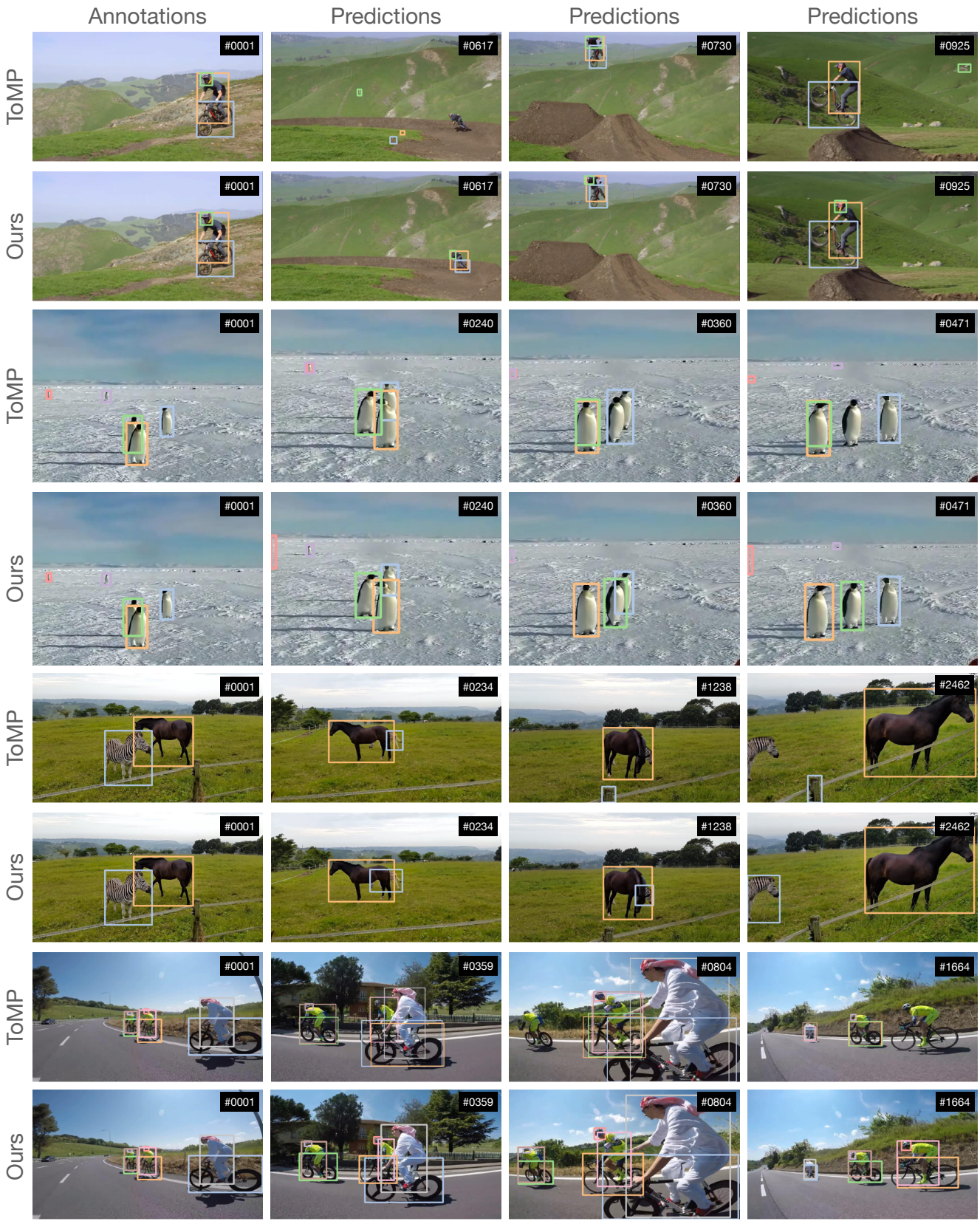


Figure 3. Visual comparison between the proposed tracker (Ours-SwinBase) and the baseline ToMP-101 on different LaGOT sequences.

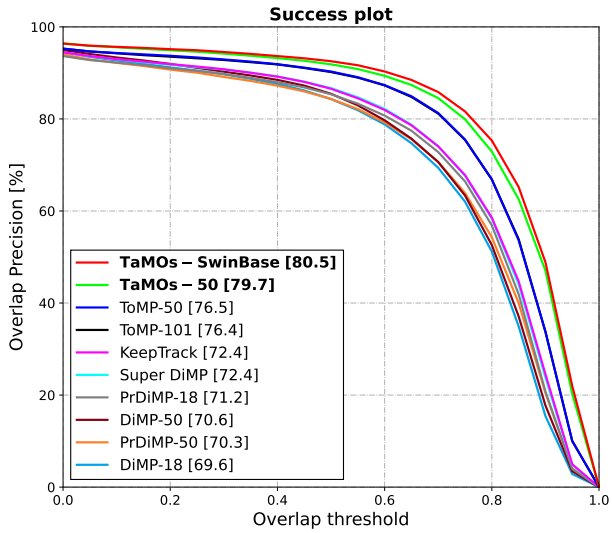


Figure 4. Success plot, showing OP_T, on ImagenetVID [27] (AUC is reported in the legend).

our tracker compared to ToMP. To summarize we observe a similar ranking between trackers on ImageNetVID and the proposed LaGOT dataset. However, LaGOT is more challenging due to the higher average track number (2.9 vs. 1.7) and the much longer sequence length (2258 vs. 312) that leads more frequently to occlusions and out-of-view events.

4. Visual Results

Visual Comparison to the State of the Art. We show visualizations of the tracking results of the baseline (ToMP-101) and our proposed tracker (TaMOs-SwinBase) on four different sequences of the proposed LaGOT benchmark in Fig. 3. The first frame specifies the target objects annotated with bounding boxes that should be tracked in the video. The other frames show predictions of both trackers. The results on the first and third sequences demonstrate that our tracker can re-detect occluded objects quickly whereas a search area based tracker is not able to re-detect the targets if they reappear outside of the search area. The second and fourth sequences show the superior robustness of our tracker. It is able to distinguish similarly looking objects better without confusing their ids. For more visual results we refer the reader to the mp4-videos submitted alongside this document. Each video shows the predictions of the proposed tracker TaMOs-SwinBase on the proposed LaGOT benchmark. Please note that we always produce a bounding box for visualization independent of its confidence score.

Failure Cases. Fig. 5 shows typical failure cases of the proposed tracker on three different sequences of the proposed LaGOT benchmark. Particularly challenging are videos that contain multiple visually similar objects since our tracker

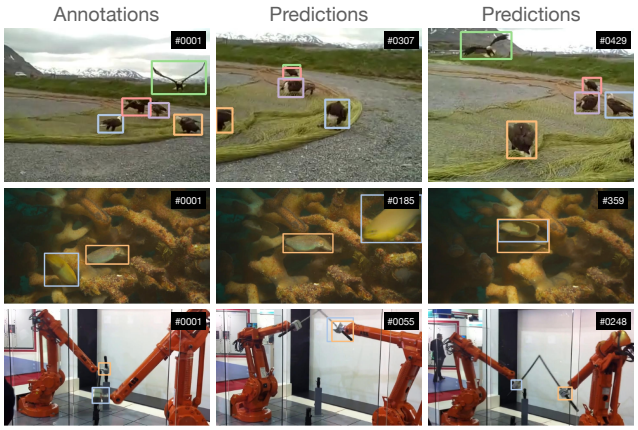


Figure 5. Visual examples of failure cases of the proposed tracker (Ours-SwinBase) on different LaGOT sequences.

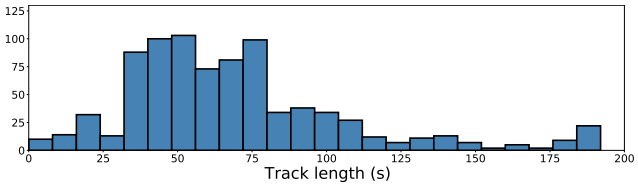


Figure 6. Track lengths distribution of the LaGOT benchmark.

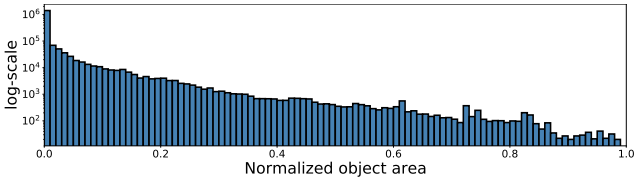


Figure 7. Object size distribution of the LaGOT benchmark.

does not employ any motion model but rather tracks the objects via the learned appearance from the first frame. Another failure case occurs when the target object is no longer visible such that our tracker might start to track a visually similar distractor instead. However, once the target reappears our globally operating tracker is usually able to re-detect it. Lastly, if multiple visually similar objects need to be tracked our tracker might fail to distinguish these objects such that it produces multiple bounding boxes with different ids for the same object.

5. Limitations and Future Work

Currently the number of objects that can be tracked is limited by the pool-size of the object embeddings. While it is possible to learn a larger pool-size it is cumbersome. Thus, an interesting direction for future research would be to generate an arbitrary number of object embedding on the fly such that any number of target objects can be tracked.

Furthermore, we propose to use an FPN to regress more

Table 5. Comparison of LaGOT and the existing datasets. Statistics is provided for test or validation set for the datasets for which test set annotations are hidden. * For MOT15-20 we report stats on the train set.

Dataset	Num Classes	Num Videos	Avg Video length (num frames)	Avg Tracks per Video	Avg Track Length (num boxes)	Avg Track Length (s)	Avg Instances per frame	Video FPS	Annotation FPS
YouTubeVOS [30]	91	474	135	1.74	27	4.5	1.64	30 FPS	6 FPS
Davis17 [25]	-	30	67	1.97	67	2.8	1.97	24 FPS	24 FPS
ImageNetVID* [9]	30	555	317	2.35	208	7	1.58	30 FPS	30 FPS
TAO* [7]	302	988	1010	5.55	21	21	3.31	30 FPS	1 FPS
BDD100k [32]	11	200	198	94.21	26	5	11.8	30 FPS	5 FPS
MOT15 [8]*	1	11	500	45.5	75	3	8	2.5-30 FPS	2.5-30 FPS
MOT16 [8]*	1	7	760	74	273	10	38	14-30 FPS	14-30 FPS
MOT20 [8]*	1	4	2233	583	572	23	150	25 FPS	25 FPS
DogThruGlasses [16]	1	30	419	3.3	352.6	11.7	2.4	30 FPS	30 FPS
GMOT-40 [1]	10	40	240	50.65	133	5.3	26.6	24-30 FPS	24-30 FPS
TrackingNet [23]	27	511	442	1	442	15	1	30 FPS	30 FPS
UAV123 [22]	8	123	915	1	915	28	1	30 FPS	30 FPS
OTB-100 [29]	16	100	590	1	590	20	1	30 FPS	30 FPS
NFS-30 [12]	15	100	479	1	479	14	1	30 FPS	30 FPS
GOT10k [15]	84	420	150	1	150	15	1	10 FPS	10 FPS
OxUvA [28]	8	200	4198	1	60	140	1	30 FPS	1 FPS
LaSOT [10]	71	280	2430	1	2430	81	1	30 FPS	30 FPS
LaGOT	102	294	2258	2.89	707	71	2.41	30 FPS	10 FPS

accurate bounding boxes for small objects and show that adding such an FPN helps. However, as in object detection, tracking extremely small objects is challenging due to the limited feature resolution when processing the full frame.

6. Datasets

Below we provide additional details about our annotated dataset, such as examples of new classes and various statistics, as well as an extensive comparison to existing datasets that focus on related tasks.

6.1. Insights

Fig. 6 shows the distribution of the track lengths in seconds for all tracks in the proposed benchmark LaGOT. We observe that most tracks are between 30 and 110 seconds long. Furthermore, Fig. 7 shows the size distribution of the annotated objects in the dataset. We conclude that various sizes are present in the dataset but large objects are rare than small ones. Further, the distribution shows that the targets are not visible in a large amount of video frames indicated by an object area of zero.

During the annotation process, we added 31 new classes: *rotor*, *fish*, *backpack*, *motor*, *wheel*, *garbage*, *drum*, *accordion*, *super-mario*, *hockey puck*, *hockey stick*, *kite-tail*, *ball*, *crown*, *stick*, *spiderweb*, *head*, *banner,face*, *bench*, *tissue-bag*, *para glider*, *star-patch*, *shadow*, *bucket*, *helicopter*, *sonic*, *hero*, *ninja-turtle*, *reflection*, *rider*.

6.2. Comparison

We provide a detailed comparison of related existing datasets in Tab. 5. We divide the table into Video Object Segmentation (VOS), Video Object Detection, Multiple Object Tracking (MOT), Generic Multiple Object Tracking (GMOT) and Single Object Tracking (SOT) datasets.

The length of VOS sequences is much shorter than in our LaGOT benchmark (2.8s/4.5s vs 71s). Similarly the video object detection dataset ImagenetVID contains shorter sequences (7s vs. 71s), fewer classes (30 vs 102) and a smaller number of average tracks per sequence (2.35 vs 2.89) than LaGOT. MOT datasets typically focus on fewer classes, contain shorter sequences or are annotated at low frame rates only. TAO contains many more classes than typical MOT datasets but provides annotations only at 1 FPS leading to a much lower average number of annotated frames per track than LaGOT (21 vs. 707). The GMOT-40 dataset contains fewer classes, fewer videos, shorter sequences and provides due to its task only annotations of one particular object class per sequence compared to LaGOT. In contrast to SOT datasets that provide only a single annotated object per sequence, LaGOT provides on average 2.89 tracks per sequence. Furthermore, it contains longer sequences than most listed SOT datasets. Overall LaGOT enables to properly evaluate the robustness and accuracy of multiple object GOT methods. A key factor are the multiple annotated tracks per sequence at a high frame rate and the relatively long sequences.

References

[1] Hexin Bai, Wensheng Cheng, Peng Chu, Juehuan Liu, Kai Zhang, and Haibin Ling. Gmot-40: A benchmark for generic multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6719–6728, June 2021. 7

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 2, 3, 4

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 4

[4] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[5] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[6] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[7] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 436–454. Springer International Publishing, 2020. 7

[8] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision (IJCV)*, 129(4):1–37, 2020. 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 7

[10] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Mingzhen Huang, Juehuan Liu, Yong Xu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision (IJCV)*, 129(2):439–461, 2021. 2, 4, 7

[11] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 4

[12] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *ICCV*, 2017. 7

[13] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 146–164, 2022. 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[15] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(5):1562–1577, 2021. 7

[16] Mingzhen Huang, Xiaoxing Li, Jun Hu, Honghong Peng, and Siwei Lyu. Tracking multiple deformable objects in egocentric videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, 2023. 7

[17] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, and Laura Leal-Taixé. Opening up open world tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19045–19055, June 2022. 3

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 1

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 2

[20] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8731–8740, June 2022. 2, 4

[21] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13444–13454, October 2021. 4

[22] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2016. 4, 7

[23] Matthias Müller, Adel Bibi, Silvio Giancola, Salman Al-Subaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 7

[24] Mubashir Noman, Wafa H Al Ghallabi, Daniya Kareem, Christoph Mayer, Akshay Dudhane, Martin Danelljan, Hisham Cholakkal, Salman Khan, Luc Van Gool, and Fahad Shahbaz Khan. Avist: A benchmark for visual object tracking in adverse visibility. In *33rd British Machine Vision Conference BMVC*, 2022. 2, 3, 4

[25] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbelaez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *CoRR*, abs/1704.00675, 2017. 7

[26] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 4, 6

[28] Jack Valmadre, Luca Bertinetto, João F. Henriques, Ran Tao, Andrea Vedaldi, Arnold W.M. Smeulders, Philip H.S. Torr, and Efstratios Gavves. Long-term tracking in the wild: a benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 7

[29] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(9):1834–1848, 2015. 7

[30] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark, 2018. 7

[31] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 341–357. Springer Nature Switzerland, 2022. 4

[32] 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 (CVPR)*, June 2020. 7

[33] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–21, 2022. 2, 4

[34] Zikun Zhou, Jianqiu Chen, Wenjie Pei, Kaige Mao, Hongpeng Wang, and Zhenyu He. Global tracking via ensemble of local trackers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8761–8770, June 2022. 4