Supplementary Material for Learning to Track with Object Permanence

Pavel Tokmakov Jie Li Wolfram Burgard Adrien Gaidon
Toyota Research Institute
first.last@tri.global

In this supplementary material, we provide additional visualizations, experimental results and implementation details that were not included in the main paper due to space limitations. We begin by describing the contents of the supplementary video, which includes qualitative examples of our algorithm's output in Section 1. In Section 2 we provide details about the datasets used in our work, and further elaborate on the PD dataset in Section 3. A real world nuScenes [5] dataset for 3D tracking could potentially be used to train our method directly, and we demonstrate preliminary results on it in Section 4. We conclude by reporting all the metrics on KITTI [6] and MOT17 [9] benchmarks in Section 5 and listing the remaining implementation details in Section 6.

1. Qualitative analysis

We demonstrate the outputs of our method on several videos from KITTI and MOT17 datasets[1]. We show raw outputs of the model, *without* any post-processing steps, such as constant velocity propagation.

00:08-00:20 In the video version of the example from Figure 5 in the main paper it is easier to see how our approach is able to correctly localize the moving car (`id 6`) occluded by the black vehicle at an intersection. Despite the complexity of the scenario (both cars, as well as the ego vehicle, are in motion) our approach successfully tracks this object as it undergoes a full occlusion.

00:24-00:49 In this example, the grey car (`id 5`) arrives at an intersection and is repeatedly occluded by three other vehicles. Again, our approach is able to maintain its trajectory throughout the whole sequence of occlusions.

00:53-01:09 The final example from the KITTI test set demonstrates occlusions by people. First, the person with `id 18` is occluded by a group of people, but his trajectory is not broken. Then the group continues forward to hide the car with `id 6`, and the person with `id 18` is re-occluded by another pedestrian, but both cases are successfully handled by our method.

01:13-01:35 In this sequence from the validation set of MOT17 it is worth paying attention to the group of people in the back on the left. They get occluded by another group, but this complex, multi-target occlusion scenario is also effectively solved by our learning-based approach.

01:39-01:50 The final positive example shows a scenario which differs from our synthetic dataset: a camera with a top-down view is flying over a street as a group of people is getting occluded by a pole. Nevertheless, our method is able to generalize to this challenging sequence.

01:54-02:05 Here we demonstrate a failure mode of our method: the car with `id 5` is occluded by a grey wagon and, for a moment, their centers overlap. CenterPoint [11] detector architecture, which serves as a basis of our model, can only predict one object center at every location. As a result, the wagon is not detected in this frame, and an id switch happens between the two trajectories. Such mistakes can often be fixed by a short-term constant velocity post-processing.

02:09-02:17 Another failure mode is shown in this example. In the complex, indoor scene shot from a person's perspective the agent with `id 45` occludes two people on the left. The furthest of them is tracked for a few frames, but is eventually lost. This is due to the fact that the person was partially occluded in the beginning, so the initial confidence of the model was low, illustrating a limitation of our approach.

2. Datasets

KITTI is a multi-object tracking benchmark capturing city driving scenarios [6]. It consists of 21 training and 29 test sequences. Cars, pedestrians, and cyclists are annotated with 2D bounding boxes at 10 FPS. Following prior work, we evaluate on the former two categories. For ablation analysis, we split each training sequence in half, and use the first half for training and the second for validation. The test set is reserved for comparison to the state of the art.

MOT17 is the standard benchmark for people tracking [9]. Unlike KITTI, most of the videos are captured with a static camera, and feature crowded indoor and outdoor areas. It consists of 7 training and 7 test sequences annotated with

[1] https://www.youtube.com/watch?v=Dj2gSJ_xILY

2D bounding boxes at 25-30 FPS. As for KITTI, we split the training videos in half to create a validation set. The standard policy on this dataset is to only report methods that do not use external data on the test set with public detections. For fairness, we compare to the state of the art on the validation set, but also report results with private detections on the test set below.

ParallelDomain (PD) is our synthetic dataset used for learning to track behind occlusions. It was collected using a state-of-the-art synthetic data generation service [3]. The dataset contains 210 photo-realistic videos with driving scenarios in city environments captured at 20 FPS. Representing crowded streets, these videos feature lots of occlusion and dis-occlusion scenarios involving people and vehicles, providing all aforementioned annotations required by our method. Each video is 10 seconds long and comes with 3 independent camera views, effectively increasing the dataset size to 630 videos. We use 582 of those for training, and the remaining 48 for validation. We ignore invisible object labels in the validation set, and evaluate all the models on visible parts of the trajectories only.

3. Statistics for the Parallel Domain Dataset

Our synthetic dataset is generated through a state-of-the-art synthetic data generation service powered by *Parallel Domain* [3]. The dataset contains 210 short snippet of crowded urban driving scenarios. Each video is 10 seconds long and is captured at 20FPS, providing 3 independent camera views. We treat the different camera views as independent videos, resulting in 630 videos in total. We split the videos into a training set with 582 videos and a validation set with 48 videos. There are no overlapping scenes between the two sets.

Each frame of a video is annotated with amodal bounding boxes defined for 9 object classes, though we focus on Pedestrians and Cars in this work. Consistent instance ids are provided across the video to support tracking association. Both visible and occluded bounding boxes are labeled with a precise visibility scores, indicating the fraction of the object which is visible in the current frame. The dataset features environments with a variable number of agents, time of day, and weather conditions. We summarize the per-class statistic regarding the length of tracks in Table 1. One can see that most trajectories span close to a half of the video, which is crucial for learning long-term tracking behaviour.

We additionally provide histograms depicting the distribution of the fraction of trajectories that are occluded within a video for the classes used in our work in Figure 1. This figure demonstrates that 64.9% of Pedestrian and 58.1% of Car trajectories are fully occluded for at least 10% of their duration, providing enough training examples for learning to track with object permanence.

The photo-realistic synthetic data along with the amodal annotations in the Parallel Domain dataset allow us to investigate a wide variety of model variants and supervision approaches with accurate annotations. As indicated in our experimental analysis, the dataset can not only be used for prototyping and analysing the proposed algorithm, but also to pre-train models with object permanence awareness that can be successfully transferred to real world datasets through our simple sim-to-real adaptation approach.

Class	# of Tracks		Avg. Length		Max Length	
	Train	Val	Train	Val	Train	Val
Pedestrian	13056	846	83.9	65.3	200	200
Car	15604	1517	105.9	94.2	200	200
Bicyclist	283	12	92.0	44.4	200	108
Bus	274	13	118.3	78.8	200	200
Caravan/RV	90	3	112.6	71.3	200	86
OtherMovable	1537	134	107.1	83.3	200	200
Motorcycle	223	24	90.1	79.1	200	192
Motorcyclist	246	28	90.2	75.0	200	200
Truck	839	76	111.2	91.4	200	200

Table 1. Parallel Domain per-category dataset statistics. Note that we count the same instances observed from different cameras separately as we treat them independently during training and evaluation.

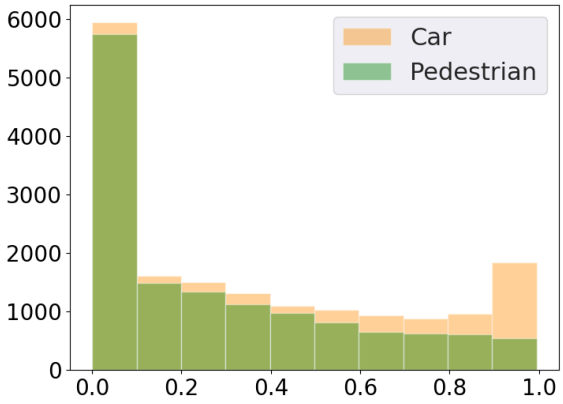

Figure 1. Histogram of occlusion ratios for the tracks in PD. We plot the fraction of the trajectory length during which the visibility score of the boxes is lower than 0.05 for both **Pedestrian** and **Car** categories. 0 indicates that the object is at least partially visible for the whole duration of the track, while 1 indicates that it is occluded for the whole duration.

4. Evaluation on nuScenes

In this section, we validate our method on the large-scale nuScenes benchmark for 3D tracking. Since this dataset is comparable in scale to PD, and provides full 3D information, we train on it directly, using the approach described in Section 3.3.2 of the main paper (see Section 6 for details). In Table 2, we compare the performance of our proposed method to the CenterTrack [10] baseline on the validation set. Our approach indeed improves the performance on the

	AMOTA	MOTA	Recall
CenterTrack [10]	6.8	6.1	0.23
PermaTrack (Ours)	**10.9**	**8.1**	**0.23**

Table 2. Validating that our method can be generalized to 3D object tracking using the validation set of nuScenes. Our method indeed outperforms the CenterTrack baseline by a significant margin on the main metrics, but a thorough investigation of 3D tracking is out of scope of this work.

main metrics. In particular, we improve the AMOTA by 4.1 points, which is a 60% relative improvement.

Note that our work focuses on 2D tracking and we only report these results to validate that our method can in principle be generalized to the 3D scenario. Tracking objects in 3D is an important problem in itself, and comes with many caveats. For instance, nuScenes is annotated at a very low frame rate (2FPS) and features a significant amount of label noise. Thoroughly investigating the effect of these factors on our method's performance is out of scope of this work.

5. Full Tables for KITTI and MOT17

In this section we report the final results of our method on KITTI and MOT17 using all the metrics on these benchmarks for reference.

KITTI uses 3 main sets of tracking metrics: HOTA-based metrics [8], CLEAR MOT metrics [4], and MT/PT/ML metrics [7]. We report them in Tables 3, 4, and 5 respectively. Full results are available on the challenge website [1].

MOT17 uses a combination of CLEAR MOT and MT/PT/ML metrics. We report all the metrics on the validation set in Table 6, and on the test set with private detections in Table 7. Full results are available on the challenge website [2].

6. Further Implementation Details

Learning to localize objects that are not visible in the current frame is challenging, and the model tends to ignore them. To avoid this, we increase the weight of the localization loss by a factor of 20 for fully occluded instances and sample sequences which contain occlusion scenarios with a probability which is proportional to the occlusion length.

For domain adaptation to KITTI [6] and MOT17 [9], we first pre-train the model on PD, and then fine-tune it jointly on PD and the corresponding dataset using the loss in Equation 2 in the main paper. The batches are sampled from each dataset with an equal probability. We use batch size 16 for all datasets, and train for 5 epochs with a learning rate of $1.25\mathrm{e}{-4}$ using the Adam optimizer. The learning rate is decreased by a factor of 10 after the 4th epoch. An epoch is defined as 5000 iterations for KITTI + PD training, and as 1600 iterations for MOT + PD due to the difference in dataset sizes.

We have found that, since videos in MOT17 are mostly captured with static cameras and the occlusions are mostly short-term, constant velocity in 2D serves as a reasonable approximation for ground truth locations of occluded people. Based on this observation, we use MOT17 sequences of length 13 during joint fine-tuning, and supervise person locations under occlusions using the pseudo-groundtruth obtained via trajectory interpolation. This strategy simplifies domain adaptation, however, as we have discussed in the main manuscript, the training set of MOT17 is too small to learn the parameters of our model from scratch.

When training on the large scale nuScenes [5] dataset we do not use PD, and instead generate pseudo-ground truth labels using the approach described in Section 3.3.2 in the main paper. For supervising 3D losses we follow all the details in [10] exactly. We have found that due to a significant amount of label noise in nuScenes using a large batch size is crucial for achieving top results. Following [10], we first pre-train our model using sequences of length 2 and batch size 64 for 70 epochs. The learning rate is set to $1.25\mathrm{e}{-4}$ and decreased by factor of 10 after 60 epochs. We then fine-tune this model using sequences of length 6 and batch size 32 for 10 epochs, decreasing the initial learning rate of $1.25\mathrm{e}{-4}$ by a factor of 10 after the 8th epoch. Finally, we freeze the backbone and further fine-tune this model with sequences of length 17 and batch size 32 to capture longer-term occlusions. This last fine-tuning stage uses the same learning rate schedule as the previous one.

References

[1] KITTI benchmark. http://www.cvlibs.net/datasets/kitti/eval_tracking_detail.php?result=82c08bddb89f9faa0fb00a60d55fea792ebede7d, March 2021. 3

[2] MOT17 benchmark. https://motchallenge.net/method/MOT=4308&chl=10, March 2021. 3

[3] Parallel domain. https://paralleldomain.com/, March 2021. 2

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 3

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 3

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012. 1, 3

[7] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene.

Category	HOTA ↑	DetA ↑	AssA ↑	DetRe ↑	DetPr ↑	AssRe ↑	AssPr ↑	LocA ↑
Car	78.0	78.3	78.4	81.7	86.5	81.1	89.5	87.1
Person	48.6	52.3	45.6	57.4	71.0	49.6	73.3	78.6

Table 3. Results of our method on the test set of the KITTI benchmark using the HOTA metrics.

Category	MOTA ↑	MOTP ↑	MODA ↑	IDSW ↓	sMOTA ↑
Car	91.3	85.7	92.1	258	78.0
Person	66.0	74.5	67.7	403	47.1

Table 4. Results of our method on the test set of the KITTI benchmark using the CLEAR MOT metrics.

Category	MT ↑	PT ↓	ML ↓	FRAG ↓
Car	85.7	11.7	2.6	250
Person	48.8	35.4	15.8	646

Table 5. Results of our method on the test set of the KITTI benchmark using the MT/PT/ML metrics.

In *2009 IEEE conference on computer vision and pattern recognition*, pages 2953–2960. IEEE, 2009. 3

[8] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, pages 1–31, 2020. 3

[9] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 3

[10] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. 2, 3, 5

[11] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1

		T.R.	IDF1 ↑	IDP ↑	IDR ↑	MOTA ↑	MOTP ↑	MT ↑	PT ↓	ML ↓	IDSW ↓	FRAG ↓
Public	PermaTrack	✗	67.0	77.5	59.0	67.8	0.178	43.7	36.3	20.1	0.8%	1.0%
	PermaTrack	✓	71.1	82.1	62.6	68.2	0.181	41.0	39.5	19.5	0.5%	1.1%
Private	PermaTrack	✗	68.2	75.9	61.9	69.4	0.18	46.3	36.0	17.7	0.9%	1.1%
	PermaTrack	✓	71.9	81.0	64.7	69.5	0.181	42.5	39.8	17.7	0.5%	1.1%

Table 6. Results of our method on the validation set of the MOT17 using private and public detections. T.R. stand for Track Rebirth post-processing from [10].

	T.R.	IDF1 ↑	IDP ↑	IDR ↑	MOTA ↑	MT ↑	ML ↓	IDSW ↓	FRAG ↓
PermaTrack	✓	68.9	75.1	63.6	73.8	43.8	17.2	3699	6132

Table 7. Results of our method on the test set of the MOT17 using private detections. This variant uses Track Rebirth (T.R.). A subset of metrics is shown, which is reported on the MOT leader-board.