Tracking the Trackers:
An Analysis of the State of the Art in Multiple Object Tracking

Laura Leal-Taixé[1,*]   Anton Milan[2,*]   Konrad Schindler[3]   Daniel Cremers[1]   Ian Reid[2]   Stefan Roth[4]

[1]Technical University Munich, Germany   [2]University of Adelaide, Australia

[3]Photogrammetry and Remote Sensing, ETH Zürich, Switzerland   [4]TU Darmstadt, Germany

Abstract

Standardized benchmarks are crucial for the majority of computer vision applications. Although leaderboards and ranking tables should not be over-claimed, benchmarks often provide the most objective measure of performance and are therefore important guides for research. We present a benchmark for Multiple Object Tracking launched in the late 2014, with the goal of creating a framework for the standardized evaluation of multiple object tracking methods. This paper collects the two releases of the benchmark made so far, and provides an in-depth analysis of almost 50 state-of-the-art trackers that were tested on over 11000 frames. We show the current trends and weaknesses of multiple people tracking methods, and provide pointers of what researchers should be focusing on to push the field forward.

1. Introduction

Evaluating and comparing multi-target tracking methods is not trivial for numerous reasons (*cf*. *e.g*. [42]). First, unlike for other tasks, such as image restoration, the ground truth, *i.e*. the perfect solution one aims to achieve, is difficult to define clearly. Partially visible, occluded, or cropped targets, reflections in mirrors or windows, and objects that very closely resemble targets all impose intrinsic ambiguities, such that even humans may not agree on one particular ideal solution. Second, a number of different evaluation metrics with free parameters and ambiguous definitions often lead to conflicting quantitative results across the literature. Finally, the lack of pre-defined test and training data makes it difficult to compare different methods in a fair way.

Even though multi-target tracking is a crucial problem in scene understanding, until recently it still lacked large-scale benchmarks to provide a fair comparison between tracking methods. Typically, methods are tuned to each individual sequence, reaching over 90% accuracy on well-known sequences like PETS [20]. Nonetheless, the real challenge for
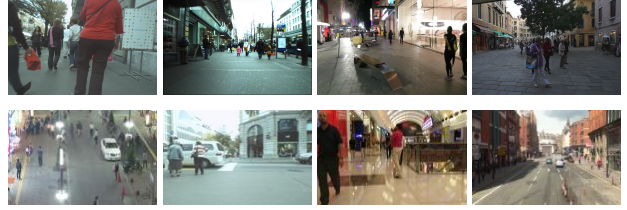

Figure 1. Examples frames of the training set for the *MOT15* release (top) and for the *MOT16* release (bottom).

a tracking system is to be able to perform well on a variety of sequences with different crowdedness levels, camera motion or illumination, ideally with a fixed set of parameters for all sequences.

In order to address this issue, we released a Multiple Object Tracking benchmark *(MOT)* in 2014. The first release of the dataset, which we refer to as *MOT15* in this paper, consists of a total of 22 sequences with 101345 annotated pedestrian bounding boxes. The second release of the dataset, referred to as *MOT16*, contains a set of 14 distinctively more crowded scenes, with 292733 annotated pedestrian boxes. A strict annotation protocol was followed to obtain accurate bounding boxes not only of pedestrians, but of a total of 12 classes of objects, including vehicles, occluders, sitting people, or motorbikes.

Since 2014, hundreds of participants have submitted their results on a standardized evaluation server, allowing for a fair comparison. In this work, we analyze 32 published trackers that have been evaluated on *MOT15* and 16 trackers on *MOT16*.[1] Having results on such a large number of sequences allows us to perform a thorough analysis on trends in tracking, currently best performing methods, and special failure cases. We aim to shed some light onto what are current weaknesses in tracking methods and therefore what to focus on for the near future in order to further improve tracking. Furthermore, we investigate the possibility of creating a 'super tracker' by selecting the best tracker for a batch of frames. Finally, we analyze the evaluation met-

---

[1]In this paper, we only consider published non-anonymous trackers that were public before 1st of March 2017. We omit LDCT [54] since it is only tested on a subset of the entire benchmark.

*denotes equal contribution.

1

Table 1. Overview of the characteristics of the data releases including both train and test splits.

Release	# Seq.	BBs	Persons	Length	Density	Tracks	HD
MOT15	22	101349	101349	16:36	8.98	1221	31.8%
MOT16	14	476532	292733	07:43	26.05	1276	85.7%

rics themselves, by conducting an experiment with human observers and analyzing the correlation of various metrics.

This paper has four main goals: *(i)* To introduce a standardized benchmark for fair evaluation of multi-target tracking methods, along with its two data releases, *MOT15* and *MOT16*; *(ii)* to analyze the performance of 32 state-of-the-art trackers on *MOT15* and 16 trackers on *MOT16*; *(iii)* to analyze common evaluation metrics using an experiment with human evaluators; *(iv)* to find the main weaknesses of current trackers and provide pointers to what the community should focus on to advance the field of multi-target tracking. The main insights gained from the analysis are: *(1)* Tracker performance is mainly influenced by the affinity metrics used, with deeply learned models giving the most encouraging results; *(2)* the expected performance of different approaches is highly correlated across videos, *i.e.* most methods perform similarly well or poorly on the same video sequence or fragment; *(3)* despite some limitations, MOTA remains the most representative measure that coincides to the highest degree with human visual assessment.

2. Related Work

Benchmarks and challenges. In the recent past, the computer vision community has developed centralized benchmarks for numerous tasks including object detection [16], pedestrian detection [13], human pose estimation [2], 3D reconstruction [53], optical flow [4, 7, 22], visual odometry [22], single-object short-term tracking [31], stereo estimation [22, 52], and video object segmentation [45], among others. Despite potential pitfalls of such benchmarks (*e.g.* [57]), they have proven to be extremely helpful in advancing the state of the art in the respective area. For multiple target tracking, in contrast, there has been very limited work on standardizing quantitative evaluation.

One of the few exceptions is the well known PETS dataset [20], targeted primarily at surveillance applications with subtasks like person counting, density estimation, flow analysis, or event recognition. Even for this widely used benchmark, we observe that tracking results are commonly obtained in an inconsistent way: using different subsets of the available data, inconsistent model training that is often prone to overfitting, varying evaluation scripts, and different detection inputs. Results are thus not easily comparable.

A well-established and useful way of organizing datasets is through standardized challenges, where results are computed in a centralized way, making comparison with any other method immediately possible. There are several

datasets organized in this fashion: Labeled Faces in the Wild [25] for unconstrained face recognition, PASCAL VOC [16] for object detection, scene classification and semantic segmentation, the DAVIS challenge [45] for video object segmentation, or the ImageNet large scale visual recognition challenge [49]. The KITTI benchmark [22] contains challenges in autonomous driving, including stereo/flow, odometry, road and lane estimation, object detection and orientation estimation, as well as tracking. Some of the sequences include crowded pedestrian crossings, making the dataset quite challenging. However, this dataset is heavily biased towards autonomous driving applications, only showing street scenes captured from a fixed camera of a driving or standing car. Cityscapes [9] is a more recent benchmark, targeting semantic and instance-level segmentation in high-definition videos, but like KITTI, is targeted for urban environments only.

With our benchmark, we want to provide a highly diverse and more challenging set of video sequences to push the limits of current multi-target tracking approaches. As can be seen in Fig. 1, our dataset includes videos from both static and moving cameras, low and high image resolution, varying weather conditions and times of the day, viewpoints, pedestrian scale, density, and more.

Analysis of the state of the art. Existing benchmarks help to push research forward and find out weaknesses of current methods for several tasks. [14] provides a thorough analysis of the state of the art in pedestrian detection on the Caltech Pedestrians dataset [13], focusing on the results of several detectors and analyzing their performance in detail. Depth estimation and optical flow evaluation were first standardized with the Middlebury dataset and the results of several methods were compared in [4]. The results from 5 years of the PASCAL VOC challenge are summarized and analyzed thoroughly in [15]. A recent study accompanying the ImageNet recognition challenge [49] also analyzes several methods that participated in that challenge, showing what tasks can still be improved and what are the main problems of current methods. The Visual Object Tracking (VOT) Challenge [32] for single object tracking, releases a yearly report with an analysis of all results presented in that year, showing the steady improvement in visual tracking. Similar to all these works, our goal is to establish a meaningful benchmark and to conduct a thorough analysis of the state of the art in multiple object tracking, providing valuable insights to the community.

3. The Multiple Object Tracking Benchmark

One of the key aspects of any benchmark is data collection. The goal of our benchmark is not only to compile yet another dataset with completely new data, but rather

to: *(i)* create a common framework to test tracking methods on; *(ii)* gather existing and new challenging sequences with very different characteristics (frame rate, pedestrian density, illumination, or point of view) in order to challenge researchers to develop more general tracking methods that can deal with all types of sequences. An overview of the characteristics of the 2015 and 2016 releases is shown in Table 1. More detailed information on each individual sequence can be found in the supplementary material.

MOT15 sequences. Our first release consisted of 22 sequences, 11 each for training and testing. The test data contains over <mark>10 minutes of footage and 61440 annotated</mark> bounding boxes, making it hard to overtune on such a large amount of data, one of the benchmark's major strengths. Among the 22 sequences, there are six new challenging high-resolution videos, four filmed from a static and two from a moving camera held at pedestrian's height. Three of them are particularly difficult: a night sequence from a moving camera and two outdoor sequences with a high density of pedestrians. The moving camera together with the low illumination creates a lot of motion blur, making this sequence extremely challenging.

MOT16 sequences. For the second data release, we focused on two aspects: *(i)* increasing the difficulty of the challenge, *e.g.* by having scenarios with a 3 times higher mean density of pedestrians; and *(ii)* improving annotations by following a strict annotation protocol that also included several classes aside from the class *pedestrian*. As can be seen in Table 1, the new data contains almost three times more bounding boxes for training and testing compared to *MOT15*. Most sequences are filmed in high resolution. Aside from pedestrians, the annotations also include other classes like vehicles, bicycles, *etc*. in order to provide contextual information for methods to exploit.

3.1. Annotation Protocol and Ground Truth

For *MOT15*, most of the ground truth bounding boxes were public. New annotations were provided for the six new sequences. One weakness of this first release was that the annotation protocol was not consistent across all sequences. In order to improve the above shortcoming, the annotations for *all MOT16* sequences have been carried out by qualified researchers from scratch following a strict protocol, and finally double-checked to ensure highest annotation accuracy. Not only pedestrians were annotated, but also vehicles, sitting people, occluding objects, as well as other significant object classes. With this fine-grained level of annotation it is possible to accurately compute the degree of occlusion and cropping of all bounding boxes, which is also provided with the benchmark. The detailed annotation protocol can be found in the supplementary material.
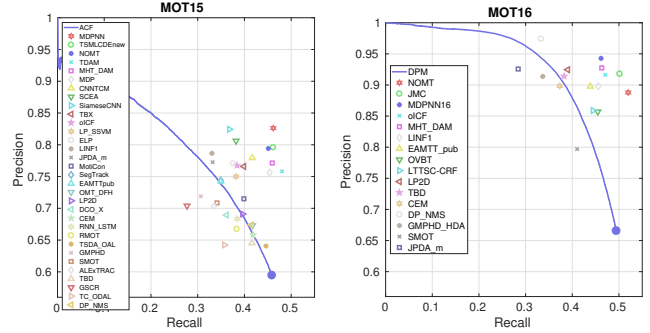

Figure 2. Precision and recall for *MOT15* (left) and *MOT16*. The accompanying detector (ACF [12] for *MOT15*, DPM [19] for *MOT16*), applied on each frame of the sequence, is plotted as a solid line. We consider all annotated pedestrians, even fully occluded ones, explaining the relatively low recall of both detectors. Trackers typically do not provide a confidence and are thus plotted with dots. Note that, when compared to the full detector set (•), most trackers are only able to improve precision, but struggle to reduce the number of missed targets.

Before the release of our benchmark, and with the exception of KITTI [22], it was common in pedestrian tracking to evaluate on a few handpicked sequences for which ground truth was known. In our setup, none of the test ground truth is public, preventing methods from overfitting to a particular scenario, with the hope to make trackers more general.

3.2. Detections

Arguably, the detector plays a crucial role for all tracking-by-detection methods. To focus the benchmark on the tracking task, we provide detections for all images. For *MOT15*, we used the Aggregated Channel Feature (ACF) detector [12], with default parameters and trained on the INRIA dataset [10]. For *MOT16*, we tested several state-of-the-art detectors, including Fast R-CNN [23]. However, when using the pre-trained, off-the-shelf model, the deformable part-based model (DPM) v5 [19] outperformed the other detectors in the task of pedestrian detection on our dataset. This is consistent with the observations made in [23], stating that the out-of-the-box R-CNN outperforms DPM in detecting all object classes except for the class *person*. All evaluated tracking methods use the same set of detections as their input: ACF detections for *MOT15* and DPM detections for *MOT16*; see Fig. 2 for precision-recall curves.

3.3. Submission and Evaluation

To limit the effect of overfitting, we follow certain submission guidelines similar to other existing benchmarks [2, 22, 49]. We limit the total number of submissions to four for one particular approach, and we enforce a minimum 72-hour gap between submissions. This has proven to be a good and effective practice to prevent participants from

overly tuning their parameters to the test set.

A critical point with any dataset is how to measure the performance of the algorithms. We will discuss and analyze different existing metrics in Sec. 6. A clear advantage of the proposed benchmark is that all metrics are computed in a centralized way and with the same exact ground truth, allowing a fair comparison of different tracking approaches.

4. Analysis of State-of-the-Art Trackers

We now present and analyze the results of all submissions, and highlight certain trends of the community. We consider all valid submissions to the benchmark that were published before March 1st, 2017 and used the provided set of detections. The total number of tracking results was 48, 32 of which were tested on *MOT15*, and the remaining 16 on *MOT16*. Note that 13 methods were tested on both datasets. Also note that a small subset of submissions[2] was done by the benchmark organizers and not by the original authors of the respective method. Results for *MOT16* are summarized in Table 2. Fig. 3 provides a graphical overview of performance as measured by MOTA for all submissions on both datasets; see the supplementary material for more details.

4.1. Trends in Tracking

We first turn our attention to the main trends in the multi-object tracking literature. Looking at Table 2, we can quickly distinguish a set of 6 top-performing trackers [8, 18, 28, 29, 50, 56], with MOTA above 40% and more than 10% Mostly Tracked trajectories. What distinguishes those trackers from the rest?

Data association. Before 2015, the community mainly focused on finding strong, preferably globally optimal, methods to solve the data association problem. The task of linking detections into a consistent set of trajectories was often cast as a graphical model and solved with k-shortest paths in DP_NMS [46], as a Linear Program solved with the simplex algorithm in LP2D [36], in a Conditional Random Field as in DCO_X [43], SegTrack [39] and LTTSC-CRF [33], or as a variational Bayesian model as in OVBT [5], to name a few. A lot of attention was also given to motion models such as SMOT [11], CEM [41], or Moti-Con [35]. The pairwise costs for matching two detections were based on either simple distances or weak appearance models. These methods achieve around 38% MOTA on *MOT16* and 25% on *MOT15*, which is 10% below current state of the art.

Affinity and appearance. More recently, the attention shifted towards building robust pairwise similarity costs,

---

[2]The methods DP_NMS, TC_ODAL, TBD, SMOT, CEM, DCO_X, LP2D were taken as baselines for the benchmark.

mostly based on strong appearance cues. This shift is clearly reflected in an increase in tracker performance, and the ability for trackers to handle more complex scenarios. The top performing methods use sparse appearance models in LINF1 [18], online appearance updates in MHT_DAM [29], integral channel feature appearance models in oICF [28], and aggregated local flow of long-term interest point trajectories in NOMT [8] to improve detection affinity. Deep learning has also had an impact on tracking, however its application to the problem at hand remains rather sparse. One example is MDPNN16 [50], which leverages Recurrent Neural Networks in order to encode appearance, motion, and interactions. JMC [56] uses deep matching to improve the affinity measure.

In summary, the main common component of top performing methods are strong affinity models. We believe this to be one of the key aspects to be addressed to further improve performance; we expect to see many more approaches that attempt to accomplish this using deep learning.

4.2. Error Analysis

We now take a closer look at the most common errors made by the tracking approaches. In Fig. 4, we show the number of false negatives (FN, blue) and false positives (FP, red) created by the trackers on average with respect to the number of FN/FP of the input detector. A ratio below 1 indicates that the trackers have improved in terms of FN/FP over the detector, while values above 1 mean a performance decrease. On the left, we see the performance for each sequence averaged over trackers; the sequences are ordered by decreasing MOTA. On the right, we show the performance for each tracker averaged over sequences.

We observe that while trackers are good at reducing FPs, most barely reduce the number of false negatives. Many of them even increase it by 20-30%. This is contrary to the common wisdom that trackers are good at filling the gaps between detections and creating full trajectories. Moreover, we can see a direct correlation between the FN and tracker performance. This is because there is a much larger number of FN than FP (detailed values can be found in the supplementary material), hence their weight on the MOTA value is much larger. One important question that arises is that if FNs are so important, why do trackers not focus more attention on reducing them?

One possible way of tackling the problem would be to get only very confident detection and create a tracker that focuses on filling the gaps to obtain long trajectories. In order to find if this could be a valid strategy, we computed the percentage of trajectories covered by the detections (determined by 50% IoU score). Taking all detections into account, as much as 18% of the trajectories are not covered by any detection. What is even more surprising is that if we drop the 10% detections with the lowest confidence, the
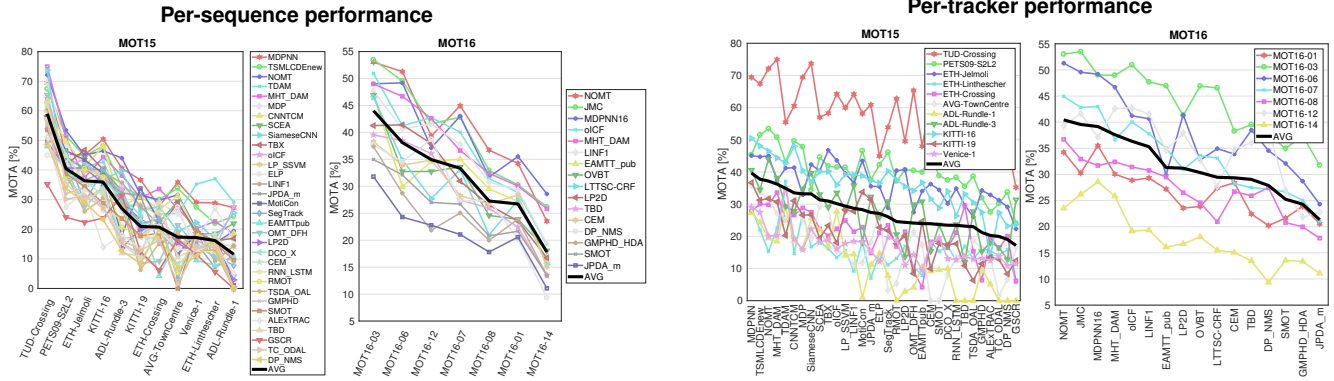
4

Figure 3. Graphical overview of all submissions. The entries are ordered from easiest sequence / best performing method, to hardest sequence / poorest performance, respectively. The mean performance across all sequences / submissions is depicted with a thick black line.
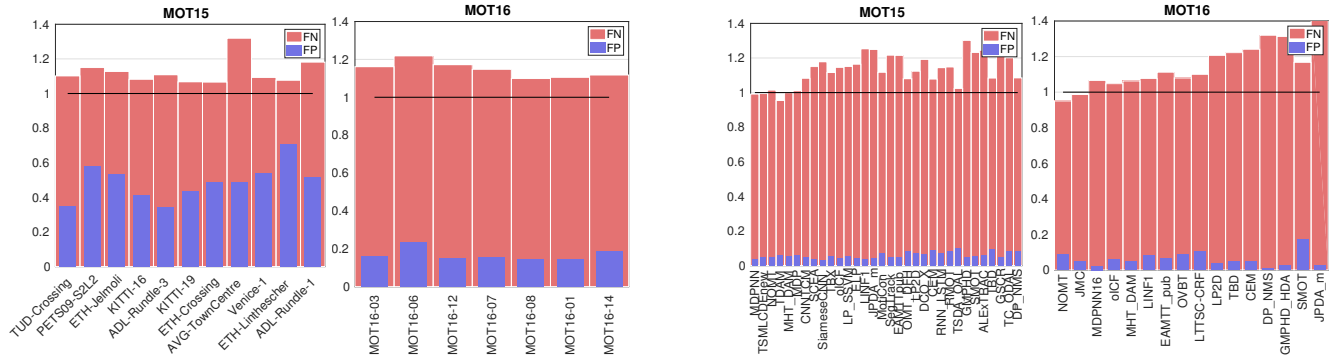

Figure 4. Detailed error analysis. The plots show the error ratios for trackers w.r.t. detector (taken at the lowest confidence threshold), for two types of errors: false positives (FP) and false negatives (FN). Values above 1 indicate a higher error count for trackers than for detectors. Note that most trackers concentrate on removing false alarms provided by the detector at the cost of eliminating a few true positives, indicated by the higher FN count.

Table 2. The *MOT16* leaderboard. Performance of several trackers according to different metrics.

Method	MOTA	MOTP	FAF	MT	ML	FP	FN	IDsw	Frag
NOMT [8]	46.4±9.9	76.6	1.6	18.3	41.4	9753	87565	359 (6.9)	504 (9.7)
JMC [56]	46.3±9.0	75.7	1.1	15.5	39.7	6373	90914	657 (13.1)	1114 (22.2)
MDPNN16 [50]	43.8±7.3	75.5	0.6	12.4	40.7	3501	98193	723 (15.7)	2036 (44.1)
oICF [28]	43.2±10.2	74.3	1.1	11.3	48.5	6651	96515	381 (8.1)	1404 (29.8)
MHT_DAM [29]	42.9±8.9	76.6	1.0	13.6	46.9	5668	97919	499 (10.8)	659 (14.2)
LINF1 [18]	41.0±9.5	74.8	1.3	11.6	51.3	7896	99224	430 (9.4)	963 (21.1)
EAMTT_pub [51]	38.8±8.5	75.1	1.4	7.9	49.1	8114	102452	965 (22.0)	1657 (37.8)
OVBT [5]	38.4±8.8	75.4	1.9	7.5	47.3	11517	99463	1321 (29.1)	2140 (47.1)
LTTSC-CRF [33]	37.6±9.9	75.9	2.0	9.6	55.2	11969	101343	481 (10.8)	1012 (22.8)
LP2D [36]	35.7±10.1	75.8	0.9	8.7	50.7	5084	111163	915 (23.4)	1264 (32.4)
TBD [21]	33.7±9.2	76.5	1.0	7.2	54.2	5804	112587	2418 (63.2)	2252 (58.9)
CEM [41]	33.2±7.9	75.8	1.2	7.8	54.4	6837	114322	642 (17.2)	731 (19.6)
DP_NMS [46]	32.2±9.8	76.4	0.2	5.4	62.1	1123	121579	972 (29.2)	944 (28.3)
GMPHD_HDA [55]	30.5±6.9	75.4	0.9	4.6	59.7	5169	120970	539 (16.0)	731 (21.7)
SMOT [11]	29.7±7.3	75.2	2.9	5.3	47.7	17426	107552	3108 (75.8)	4483 (109.3)
JPDA_m [48]	26.2±6.1	76.3	0.6	4.1	67.5	3689	130549	365 (12.9)	638 (22.5)

number of ==completely uncovered tracks goes up to 55%.== Of course, these trajectories will never be recovered by any

tracker because there is no remaining evidence at all for that pedestrian. Thereby, this strategy would lead to more FNs

5

Table 3. *MOT16* trackers and their characteristics.

Method	Box-box Affinity	Appearance	Optimization	Extra Inputs	Online
NOMT [8]	Interest Point Trajectories	✓	CRF	Optical flow	✗
JMC [56]	DeepMatching	✓	Multicut	Non-NMS dets	✗
MDPNN16 [50]	RNN (motion, appearance, interactions)	✓	Markov Decision Process	—	✓
oICF [28]	Motion model + MIL on appearance	✓	Kalman filter	—	✓
MHT_DAM [29]	Regression classifier appearance	✓	Multiple Hypothesis	—	✗
LINF1 [18]	Sparse representations appearance	✓	MCMC	—	✗
EAMTTpub [51]	2D distances	✗	Particle Filter	Non-NMS dets	✓
OVBT [5]	Dynamics from flow	✓	Variational EM	Optical flow	✓
LTTSC-CRF [33]	SURF	✓	CRF	SURF	✗
LP2D [36]	2D image distances, IoU	✗	Global, LP	—	✗
TBD [21]	IoU + NCC	✓	Hungarian algorithm	—	✗
CEM [41]	2D velocity difference	✗	L-BFGS + greedy sampling	—	✗
DP_NMS [46]	2D image distances	✗	k-shortest paths	—	✗
GMPHD_HDA [55]	HoG similarity, color histogram	✓	Gaussian mixture PHD filter	HoG	✓
SMOT [11]	Target dynamics	✗	Hankel Total Least Squares	—	✗
JPDA_m [48]	Mahalanobis distance	✗	LP	—	✗

in the end than current strategies that focus on better recovering the feasible trajectories.

5. Predicting Performance

In this section, we investigate the question whether it is possible to predict how currently available tracking methods would perform on a particular video. This would allow us to create a 'super tracker' by choosing the best approach each time we are confronted with a new video. We carry out this analysis at two granularity levels: *(a)* at a sequence level, and *(b)* at the level of short video snippets. For the latter, the entire dataset is divided into temporally overlapping 50-frame long fragments, with a stride of 25 frames. The benefit of this fine-grained analysis is twofold. On one hand, it allows us to precisely pinpoint the strengths and weaknesses of current methods for a particular situation without averaging over the entire video sequence. On the other hand, this approach generates enough data for training a classifier.

Fine-grained performance analysis. Results of this fine-grained evaluation are illustrated in Figs. 12 and 6. Fig. 12 depicts an example of this evaluation on the KITTI-19 sequence. We only show a subset of all trackers to maintain readability. Please refer to the supplemental material for more examples. Fig. 6 shows the same evaluation for *all* trackers and *all* sequences as a heatmap. On the right, three example frames for the locally highest, locally lowest, and locally intermediate MOTA averaged across all submissions are shown. It is interesting to note that for both *MOT15* and *MOT16*, two out of the three examples are picked from the same sequence, which demonstrates the performance difference within one video sequence. Looking at the two hardest examples, we can observe two similarities. First, they both fall within the 'driving' scene scenario, which is typically more difficult due to camera motion. Second, these fragments contain relatively few true targets such that
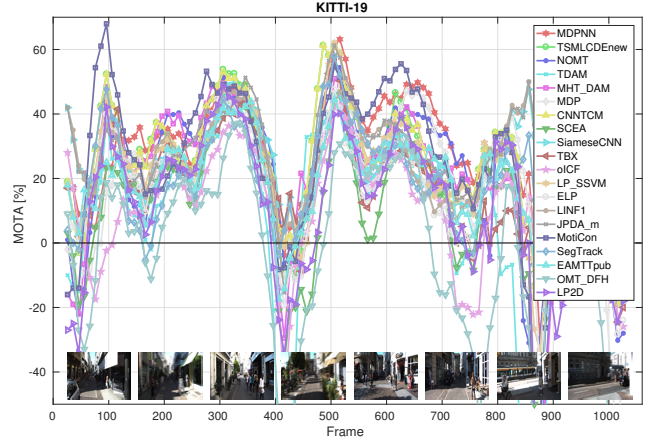

Figure 5. An example of fine-grained analysis for the KITTI-19 sequence. Each line represents the *local* performance of a tracker, measured by MOTA within a 50-frame (roughly 3 seconds long) segment. Note the extreme within-sequence variation.

the MOTA is largely dominated by the false positive tracks which are not suppressed. Please see the supplemental video for this and more examples.

For both levels of granularity we performed the following experiment. We took each sequence (or a shorter video fragment), computed several features and trained a linear multi-class SVM using cross validation to predict which of the top-three trackers produces the best result on that sequence. We used 7 different features: min/max scale ratio

Table 4. Combining tracker output. The results reflect MOTA of the state-of-the-art method (best) and a modified version by considering the top-3 performing submissions. See text for details.

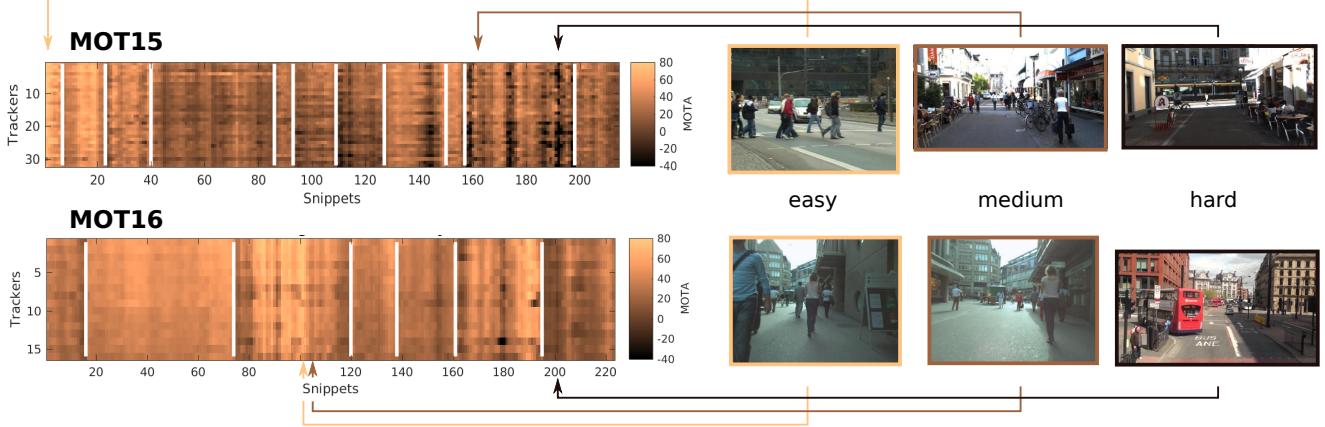| Dataset	per-sequence			fine-grained	
	best	predict	oracle	predict	oracle
MOT15	36.6	36.7	40.5	39.5	42.1
MOT16	46.4	46.2	47.4	46.0	48.3

6

Figure 6. A fine-grained video complexity analysis for *MOT15* (top) and *MOT16* (bottom) for all trackers. Each row shows a performance heatmap for all trackers, measured by MOTA in 50-frame snippets. The three frames on the right represent the easiest, average, and most difficult video section, averaged over all trackers. Note the high within sequence performance difference in both cases.

of the top 50% detections, the mean number of detections per frame, the mean detection confidence, the mean mean and mean max flow magnitudes of entire frames, and those of only the non-person regions. The first three features represent scene geometry and saliency, and indicate whether people are easy to be detected. The optic flow features estimate the presence of camera motion.

The prediction for each temporal segment was made using the split that did not contain this data point in its training set. To obtain the final tracking result, the individual outputs are simply concatenated without sophisticated association schemes. We also provide an upper bound on the best possible performance gain by selecting the optimal tracker result for each video segment.

The results are summarized in Tab. 4. Interestingly, even with the oracle prediction for each segment, the overall improvement on MOTA compared to the currently best performing method is moderate. It is 5.5 percentage points for *MOT15* and merely 2.9 percentage points for *MOT16*. Using the trained SVMs prediction, we can only slightly improve on *MOT15*, however the classifier is too error-prone on the *MOT16* dataset, even leading to a minor decay. Considering the complexity analysis in Fig. 6, this is not entirely surprising. *MOT15* contains a much higher diversity of results for a particular video fragment, while *MOT16* is fairly homogeneous in comparison. This makes it hard to predict the best method, but also to improve the overall performance using the oracle prediction.

6. Analysis of the Evaluation Metrics

One of the key aspects of any benchmark is the evaluation protocol. In the case of multi-object tracking, the CLEAR metrics [27] have emerged as one of the standards. By measuring the intersection over union of tracker bounding boxes and matched ground truth annotations, Accuracy

(MOTA) and Precision (MOTP) can be computed. Precision measures how well the persons are localized, while Accuracy evaluates how many distinct errors such as missed targets (FN), ghost tracks (FP), or identity switches (IDSW) are made. Another set of measures that is widely used is that of [37]: mostly tracked (MT), mostly lost (ML), and partially tracked (PT) pedestrians. These numbers give a very good intuition on the overall performance of the method.

6.1. Do Metrics and Humans Agree?

Our first goal is to find out how well these metrics reflect the perception of a human evaluator on the quality of a tracking result. To that end, we perform the following study. We create a visual quality assessment task by asking each participant to choose the best tracker among two randomly sampled trackers from the list. The participants only see the video results of the two trackers playing at the same time, and have the possibility to go forward and backward or stop the video at any point. In total, we collected results from 500 participants, both from the vision community as well as external visitors of our benchmark website.

Each tracker pair is then ranked according to each measure. In Fig. 7 (top), we plot the votes that agreed with that ranking with respect to the difference in the metric. We can see that when the two trackers have a difference of 5 percentage points in MOTA or less, the human visual assessment cannot distinguish which one is better, since the votes are split almost 50/50. The same observation applies to the Mostly Tracked (MT) metric as well as Recall (Rcll). After that, the opinions quickly align with the metrics and it is easy for a human observer to distinguish the better tracker. We thus conclude that all trackers with a difference of 5 percentage points or less in MOTA, MT, or Recall can be considered to have a largely equivalent performance.

Another interesting question that we aim to answer with this experiment is which measure reflects the human vi-
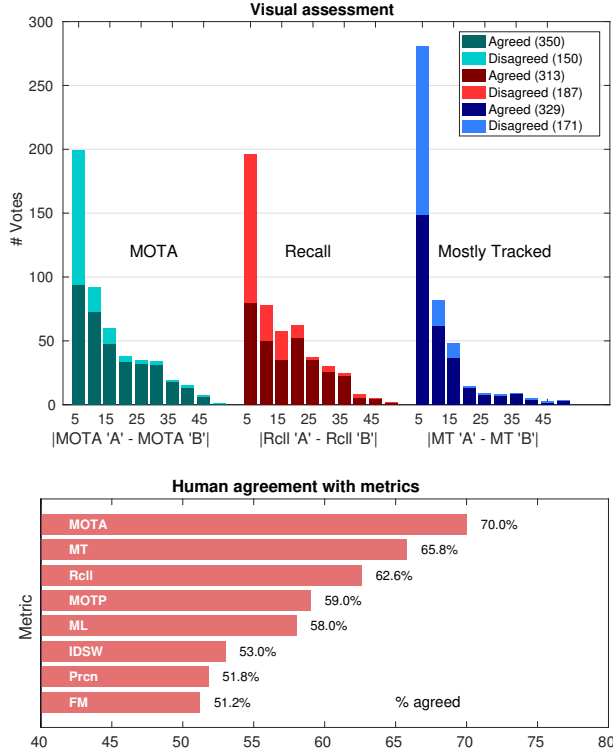
Figure 7. Results for the tracker visual quality assessment. *Top:* Votes from human observers that agree with the metrics (MOTA, Recall and Mostly Tracked) vs the performance difference between tracker A and B under evaluation. *Bottom:* Percentage of votes that align with each of the performance metrics.

sual assessment best. In Fig. 7 (bottom) we plot the percentage of votes that agree with the metric's assessment. While often highly criticized, MOTA is still the measure that best aligns with the human visual assessment. Mostly Tracked (MT) follows as second-best measure. Unsurprisingly, identity switches do not have much of an impact on the quality assessment. This reflects that human observers give much more importance to the fact that people are detected rather than them being correctly tracked.

6.2. Are All Metrics Necessary?

While it is clear that Accuracy (MOTA) and Precision (MOTP) measure two different aspects in tracking, and MOTP is depending mostly on the detector's ability to localize bounding boxes, it is often unclear whether Mostly Tracked (MT) and MOTA actually measure the same characteristics of a tracker. In Fig. 8, we plot the correlation of several evaluation metrics. Each point represents a result of a tracker in a snippet (*cf*. Sec. 5). Each color belongs to one test sequence of *MOT16*. As expected, Recall and MOTA are highly correlated. Recalling the figures from Tab. 2, we see that the number of missed targets (FN) is typically two to three orders of magnitude higher than FP and ID, which are the other two components of MOTA. Interestingly, even
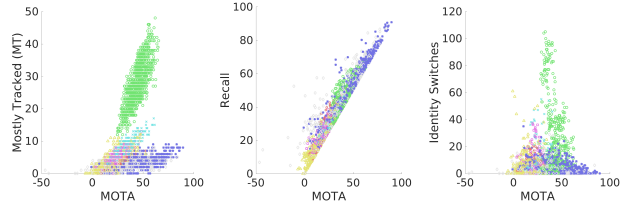

Figure 8. Metric correlations.

though both measures are strongly correlated, Fig. 7 suggests that humans agree more with the MOTA measure than with the Recall when asked to visually judge a tracker.

MOTA and MT, on the other hand, do not have such a distinct linear correlation. The points are clearly clustered by sequence, which might indicate that one of the measures actually depends on the scene under evaluation. The trend in the community is to report both measures, which is also supported by this experiment. As expected, identity switches and MOTA are hardly correlated at all.

6.3. Limitations of the Experiment

Assessing the performance of a multi-object tracking method can vary significantly depending on the application at hand. In surveillance scenarios, it is typically more important to have no false negatives so that no person is missed, while a few false alarms can be easily controlled by humans in the loop. For applications such as crowd analysis or people motion analysis, it is desirable to have a few very reliable trajectories so that trajectory analysis is accurate, thereby, false positives are not desired. In sports, it is crucial to maintain the players' identities robustly to obtain the most reliable statistics. When creating a benchmark for all-purpose multiple people tracking, it is less clear which metrics to focus on. In absence of a concrete application, human judgement is a useful point of reference.

Of course, the proposed experiment has its limitations. Results in crowded scenes consist of several tens of bounding boxes following numerous pedestrians of different sizes. Studies [1, 47] have shown that humans are able to track an average of 4 objects at normal speed, reaching up to 8 objects as very slow speeds but are limited to possibly a single object at very high speeds. Even though the human observer can stop/restart the video at any time, the complexity of the task remains very high for a human observer. It is therefore likely that he/she will miss small differences between two similarly performing trackers.

Another issue is regarding the relatively low importance that observers give to identity switches when judging a tracker. One reason for this could also be due to visualization. Each image is overlaid with the bounding box of the detected person and the color of the bounding box represent its identity. While a false positive can be easily spotted by seeing bounding boxes that are overlaid to background rather than person (similar for false negatives), an identity

switch is just shown as a change of color in the bounding box, possibly after a long gap, which might not be perceived by the human observer. Despite the aforementioned limitations, the experiment offers valuable insights that match the community's perception of existing performance metrics.

7. Conclusion and Future Work

We have introduced a standardized benchmark for fair evaluation of multi-target tracking methods and its two data releases with circa 23000 frames of footage and almost 400000 annotated pedestrians. We have analyzed the commonly used evaluation metrics with an experiment with human evaluators, and found that even though MOTA have been frequently criticized by the community, it still remains the most representative measure that coincides to the highest degree with human visual assessment. We have further analyzed the performance of 32 state-of-the-art trackers on *MOT15* and 16 trackers on *MOT16*, obtaining several insights. In particular, that trackers performance is very influenced by the affinity metrics used, and deeply learned models are currently giving the most encouraging results. Furthermore, we found the expected performance of different approaches highly correlated across videos, *i.e.* most methods perform similarly well or poorly on the same video sequence or fragment. We believe that our Multiple Object Tracking Benchmark and the presented systematic analysis of existing tracking algorithms helps to identify their strengths and weaknesses and paves the way for future innovations.

References

[1] G. A. Alvarez and S. L. Franconeri. Journal of vision. *How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism*, 7(14), 2007. 8

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR 2014*, pages 3686–3693, June 2014. 2, 3

[3] S.-H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *CVPR 2014*. 18

[4] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1):1–31, Mar. 2011. 2

[5] Y. Ban, S. Ba, X. Alameda-Pineda, and R. Horaud. Tracking multiple persons based on a variational bayesian model. *ECCV Workshops - Benchmarking Multi-Target Tracking*, 2016. 4, 5, 6

[6] A. Bewley, L. Ott, F. Ramos, and B. Upcroft. Alextrac: Affinity learning by exploring temporal reinforcement within association chains. *ICRA*, 2016. 18

[7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV 2012*, volume 4, pages 611–625, Oct. 2012. 2

[8] W. Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *ICCV 2015*. 4, 5, 6, 18

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR 2016*. 2

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR 2005*, pages 886–893. 3

[11] C. Dicle, M. Sznaier, and O. Camps. The way they move: Tracking multiple targets with similar appearance. In *ICCV 2013*. 4, 5, 6, 18

[12] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 36(8):1532–1545, 2014. 3

[13] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR 2009*. 2

[14] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34, 2012. 2

[15] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 2

[16] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. 2012. 2

[17] L. Fagot-Bouquet, R. Audigier, Y. Dhome, and F. Lerasle. Online multi-person tracking based on global sparse collaborative representations. *ICIP*, 2015. 18

[18] L. Fagot-Bouquet, R. Audigier, Y. Dhome, and F. Lerasle. Improving multi-frame data association with sparse representations for robust near-online multi-object tracking. *ECCV*, 2016. 4, 5, 6, 18

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010. 3

[20] J. Ferryman and A. Ellis. PETS2010: Dataset and challenge. In *Advanced Video and Signal Based Surveillance (AVSS)*, 2010. 1, 2

[21] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *PAMI*, 2014. 5, 6, 18

[22] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI Vision Benchmark Suite. In *CVPR 2012*. 2, 3

[23] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. *CVPR*, 2015. 3

[24] R. Henschel, L. Leal-Taixé, B. Rosenhahn, and K. Schindler. Tracking with multi-level features. *arXiv:1607.07304*, 2016. 18

[25] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachussetts, Amherst, 2007. 2

[26] J. Ju, D. Kim, B. Ku, D. Han, and H. Ko. Online multi-object tracking with efficient track drift and fragmentation handling. *Int. J. Opt. Soc. Am. A*, 2017. 18

[27] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation for face, text and vehicle detection and tracking in video: data, metrics, and protocol. *PAMI*, 31(2), 2009. 7

[28] H. Kieritz, S. Becker, W. Hübner, and M. Arens. Online multi-person tracking using integral channel features. *AVSS*, 2016. 4, 5, 6, 18

[29] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg. Multiple hypothesis tracking revisited: Blending in modern appearance model. In *ICCV 2015*. 4, 5, 6, 18

[30] H. Ko. Online multi-person tracking with two-stage data association and online appearance model learning. *IET Computer Vision*, 2017. 18

[31] M. Kristan et al. The visual object tracking VOT2014 challenge results. In *European Conference on Computer Vision Workshops (ECCVW). Visual Object Tracking Challenge Workshop*, 2014. 2

[32] M. Kristan et al. The visual object tracking VOT2016 challenge results. In G. Hua and H. Jégou, editors, *ECCV 2016 Workshops*, pages 777–823. 2016. 2

[33] N. Le, A. Heili, and M. Odobez. Long-term time-sensitive costs for CRF-based tracking by detection. *ECCV Workshops - Benchmarking Multi-Target Tracking*, 2016. 4, 5, 6

[34] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler. Learning by tracking: Siamese CNN for robust target association. *CVPR Workshops - DeepVision*, 2016. 18

[35] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese. Lerning an image-based motion context for multiple people tracking. *CVPR*, 2014. 4, 18

[36] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. *ICCV. 1st Workshop on Modeling, Simulation and Visual Analysis of Large Crowds*, 2011. 4, 5, 6, 18

[37] Y. Li, C. Huang, and R. Nevatia. Learning to associate: HybridBoosted multi-target tracker for crowded scene. In *CVPR 2009*. 7

[38] N. McLaughlin, J. M. D. Rincon, and P. Miller. Enhancing linear programming with motion modeling for multi-target tracking. *WACV*, 2015. 18

[39] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid. Joint tracking and segmentation of multiple targets. In *CVPR 2015*. 4, 18

[40] A. Milan, S. Rezatofighi, A. Dick, I. Reid, and K. Schindler. Online multi-target tracking using recurrent neural networks. *AAAI*, 2017. 18

[41] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *PAMI*, 36(1):58–72, 2014. 4, 5, 6, 18

[42] A. Milan, K. Schindler, and S. Roth. Challenges of ground truth evaluation of multi-target tracking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 735–742, June 2013. 1

[43] A. Milan, K. Schindler, and S. Roth. Multi-target tracking by discrete-continuous energy minimization. *TPAMI*, 2016. 4, 18

[44] Y. Min and J. Yunde. Temporal dynamic appearance modeling for online multi-person tracking. *arXiv:1510.02906*, 2015. 18

[45] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. *CVPR*, 2016. 2

[46] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR 2011*. 4, 5, 6, 18

[47] Z. W. Pylyshyn and R. W. Storm. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, 3(3), 1988. 8

[48] S. H. Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid. Joint probabilistic data association revisited. In *ICCV 2015*. 5, 6, 18

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. *ImageNet Large Scale Visual Recognition Challenge*. 2014. 2, 3

[50] A. Sadeghian, A. Alahi, and S. Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. *arXiv:1701.01909*, 2017. 4, 5, 6, 18

[51] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro. Multi-target tracking with strong and weak detections. *ECCV Workshops - Benchmarking Multi-Target Tracking*, 2016. 5, 6, 18

[52] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, Apr. 2002. 2

[53] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR 2006*, pages 519–528. 2

[54] F. Solera, S. Calderara, and R. Cucchiara. Learning to divide and conquer for online multi-target tracking. *ICCV*, 2015. 1

[55] Y. Song and M. Jeon. Online multiple object tracking with the hierarchically adopted gm-phd filter using motion and appearance. *ICCE*, 2016. 5, 6, 18

[56] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Multi-person tracking by multicuts and deep matching. *ECCV Workshops - Benchmarking Multi-Target Tracking*, 2016. 4, 5, 6

[57] A. Torralba and A. Efros. Unbiased look at dataset bias. In *CVPR 2011*. 2

[58] B. Wang, K. L. Chan, L. Wang, B. Shuai, Z. Zuo, T. Liu, and G. Wang. Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association. *CVPR Workshops, DeepVision*, 2016. 18

10

[59] B. Wang, G. Wang, K. L. Chan, and L. Wang. Tracklet association by online target-specific metric learning and coherent dynamics estimation. *arXiv:1511.06654*, 2015. 18

[60] S. Wang and C. Fowlkes. Learning optimal parameters for multi-target tracking with contextual interactions. *IJCV*, 2016. 18

[61] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *International Conference on Computer Vision (ICCV)*, pages 4705–4713, 2015. 18

[62] J. Yoon, C. Lee, M. Yang, and K. Yoon. Online multi-object tracking via structural constraint event aggregation. *CVPR*, 2016. 18

[63] J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon. Bayesian multi-object tracking using motion context from multiple objects. In *WACV*, 2015. 18

Supplementary Material

The following sections contain additional material that was not included in the main paper. In particular, we first provide a more thorough description of the two datasets and the annotation rules for *MOT16*. We also include full leaderboards for both releases listing all metrics of the 48 analyzed trackers. Finally, we provide additional figures for the error analysis and the fine-grained analysis.

The Multi-Object Tracking Benchmark

MOT15 includes a total of 22 sequences, of which we use half for training and half for testing. Most of these sequences had been previously introduced and used by the multi-target tracking community. For *MOT15* we rely on the publicly available ground truth. In Table 5, we detail the characteristics of each of the sequences including length of the sequence, number of pedestrian tracks, number of pedestrian bounding boxes, density of the scene, moving or static camera, viewpoint, and weather conditions.

In contrast to the 2015 edition, *MOT16* consists almost exclusively of novel, high-definition videos, *all* of which have been (re-)annotated following a consistent protocol (see Section 7). *MOT16* includes a total of 14 sequences, of which we use half for training and half for testing. For these sequences, we also provide further information in Table 6. In Table 7, we list the types of annotations that we provide with the dataset, including cars, motorbikes, and bicycles in addition to pedestrians.

Annotation Rules

To mitigate the effect of poor and inconsistent labeling, for *MOT16*, we follow a set of rules to annotate every moving person or vehicle within each sequence with a bounding box as accurately as possible. In the following we define a clear protocol that was obeyed throughout the entire dataset to guarantee consistency.

Target class

In this benchmark we are interested in tracking moving objects in videos. In particular, we are interested in evaluating multiple people tracking algorithms, therefore, people will be the center of attention of our annotations. We divide the pertinent classes into three categories:

1. *moving* or *standing* pedestrians;

2. people that are *not in an upright position* or artificial representations of humans; and

3. *vehicles* and *occluders*.

In the first group, we annotate all moving or standing (upright) pedestrians that appear in the field of view and can be determined as such by the viewer. People on bikes or skateboards will also be annotated in this category (and are typically found by modern pedestrian detectors). Furthermore, if a person *briefly* bends over or squats, *e.g.* to pick something up or to talk to a child, they shall remain in the standard *pedestrian* class. The algorithms that submit to our benchmark are expected to track these targets.

In the second group we include all people-like objects whose exact classification is ambiguous and can vary depending on the viewer, the application at hand, or other factors. We annotate all static people that are not in an upright position, *e.g.* sitting, lying down. We also include in this category any artificial representation of a human that might cause a strong detection response, such as mannequins, pictures, or reflections. People behind glass should also be marked as distractors. The idea is to use these annotations in the evaluation such that an algorithm is neither penalized nor rewarded for tracking, *e.g.*, a sitting person or a reflection.

In the third group, we annotate all moving vehicles such as cars, bicycles, motorbikes, and non-motorized vehicles (*e.g.*, strollers), as well as other potential occluders. We will not evaluate specifically against these annotations, rather they are provided to the users both for training purposes and for computing the level of occlusion of pedestrians. Static vehicles (parked cars, bicycles) are not annotated as long as they do not occlude any pedestrians.

The rules are summarized in Tab. 8 and in Fig. 9 we present a diagram of the classes of objects we annotate, as well as a sample frame with annotations.

Bounding box alignment

The bounding box is aligned with the object's extent as accurately as possible. The bounding box should contain all pixels belonging to that object and at the same time be as tight as possible, *i.e.* no pixels should be left outside the box. This means that a walking side-view pedestrian will typically have a box whose width varies periodically with the stride, while a front view or a standing person will maintain a more constant aspect ratio over time. If the person is partially occluded, the extent is estimated based on other available information such as expected size, shadows, reflections, previous and future frames, and other cues. If a person is cropped by the image border, the box is estimated beyond the original frame to represent the entire person and to estimate the level of cropping. If an occluding object cannot be accurately enclosed in one box (*e.g.*, a tree with branches or an escalator may require a large bounding box where most of the area does not belong to the actual ob-

Training sequences									
Name	FPS	Resolution	Length	Tracks	Boxes	Density	Camera	Viewpoint	Conditions
TUD-Stadtmitte	25	640x480	179 (00:07)	10	1156	6.5	static	medium	normal
TUD-Campus	25	640x480	71 (00:03)	8	359	5.1	static	medium	normal
PETS09-S2L1	7	768x576	795 (01:54)	19	4476	5.6	static	high	normal
ETH-Bahnhof	14	640x480	1000 (01:11)	171	5415	5.4	moving	low	normal
ETH-Sunnyday	14	640x480	354 (00:25)	30	1858	5.2	moving	low	shadows
ETH-Pedcross2	14	640x480	840 (01:00)	133	6263	7.5	moving	low	shadows
ADL-Rundle-6	30	1920x1080	525 (00:18)	24	5009	9.5	static	low	indoor
ADL-Rundle-8	30	1920x1080	654 (00:22)	28	6783	10.4	moving	medium	night
KITTI-13	10	1242x375	340 (00:34)	42	762	2.2	moving	medium	shadows
KITTI-17	10	1242x370	145 (00:15)	9	683	4.7	static	medium	shadows
Venice-2	30	1920x1080	600 (00:20)	26	7141	11.9	static	medium	normal
Total training			**5503 (06:29)**	**500**	**39905**	**7.3**			

Testing sequences									
Name	FPS	Resolution	Length	Tracks	Boxes	Density	Camera	Viewpoint	Conditions
TUD-Crossing	25	640x480	201 (00:08)	13	1102	5.5	static	medium	normal
PETS09-S2L2	7	768x576	436 (01:02)	42	9641	22.1	static	high	normal
ETH-Jelmoli	14	640x480	440 (00:31)	45	2537	5.8	moving	low	shadows
ETH-Linthescher	14	640x480	1194 (01:25)	197	8930	7.5	moving	low	shadows
ETH-Crossing	14	640x480	219 (00:16)	26	1003	4.6	moving	low	normal
AVG-TownCentre	2.5	1920x1080	450 (03:45)	226	7148	15.9	static	high	normal
ADL-Rundle-1	30	1920x1080	500 (00:17)	32	9306	18.6	moving	medium	normal
ADL-Rundle-3	30	1920x1080	625 (00:21)	44	10166	16.3	static	medium	shadows
KITTI-16	10	1242x370	209 (00:21)	17	1701	8.1	static	medium	shadows
KITTI-19	10	1242x374	1059 (01:46)	62	5343	5.0	moving	medium	shadows
Venice-1	30	1920x1080	450 (00:15)	17	4563	10.1	static	medium	normal
Total testing			**5783 (10:07)**	**721**	**61440**	**10.6**			

Table 5. Overview of the sequences included in the *MOT15* release.

Training sequences									
Name	FPS	Resolution	Length	Tracks	Boxes	Density	Camera	Viewpoint	Conditions
02	30	1920x1080	600 (00:20)	49	17,833	29.7	static	medium	cloudy
04	30	1920x1080	1,050 (00:35)	80	47,557	45.3	static	high	night
05	14	640x480	837 (01:00)	124	6,818	8.1	moving	medium	sunny
09	30	1920x1080	525 (00:18)	25	5,257	10.0	static	low	indoor
10	30	1920x1080	654 (00:22)	54	12,318	18.8	moving	medium	night
11	30	1920x1080	900 (00:30)	67	9,174	10.2	moving	medium	indoor
13	25	1920x1080	750 (00:30)	68	11,450	15.3	moving	high	sunny
Total training			**5,316 (03:35)**	**512**	**110,407**	**20.8**			

Testing sequences									
Name	FPS	Resolution	Length	Tracks	Boxes	Density	Camera	Viewpoint	Conditions
01	30	1920x1080	450 (00:15)	23	6,395	14.2	static	medium	cloudy
03	30	1920x1080	1,500 (00:50)	148	104,556	69.7	static	high	night
06	14	640x480	1,194 (01:25)	217	11,538	9.7	moving	medium	sunny
07	30	1920x1080	500 (00:17)	55	16,322	32.6	moving	medium	shadow
08	30	1920x1080	625 (00:21)	63	16,737	26.8	static	medium	sunny
12	30	1920x1080	900 (00:30)	94	8,295	9.2	moving	medium	indoor
14	25	1920x1080	750 (00:30)	230	18,483	24.6	moving	high	sunny
Total testing			**5,919 (04:08)**	**830**	**182,326**	**30.8**			

Table 6. Overview of the sequences currently included in the *MOT16* benchmark.

Annotation classes												
Sequence	Pedestrian	Person on vehicle	Car	Bicycle	Motorbike	Non-motorized vehicle	Static person	Distractor	Occluder on ground	Occluder full	Reflection	Total
01	6,395	346	0	341	0	0	4,790	900	3,150	0	0	15,922
02	17,833	1,549	0	1,559	0	0	5,271	1,200	1,781	0	0	29,193
03	104,556	70	1,500	12,060	1,500	0	6,000	0	24,000	13,500	0	163,186
04	47,557	0	1,050	11,550	1,050	0	4,798	0	23,100	18,900	0	108,005
05	6,818	315	196	315	0	11	0	16	0	0	0	7,671
06	11,538	150	0	118	0	0	269	238	109	0	0	12,422
07	16,322	0	0	0	0	0	2,023	0	1,920	0	0	20,265
08	16,737	0	0	0	0	0	1,715	2,719	6,875	0	0	28,046
09	5,257	0	0	0	0	0	0	1,575	1,050	0	948	8,830
10	12,318	0	25	0	0	0	1,376	470	2,740	0	0	16,929
11	9,174	0	0	0	0	0	0	306	596	0	0	10,076
12	8,295	0	0	0	0	0	1,012	765	1,394	0	0	11,464
13	11,450	0	4,484	103	0	0	0	4	2,542	680	0	19,263
14	18,483	0	1,563	0	0	0	712	47	4,062	393	0	25,260
Total	292,733	2,430	8,818	26,046	2,550	11	27,966	8,238	73,319	33,473	948	476,532

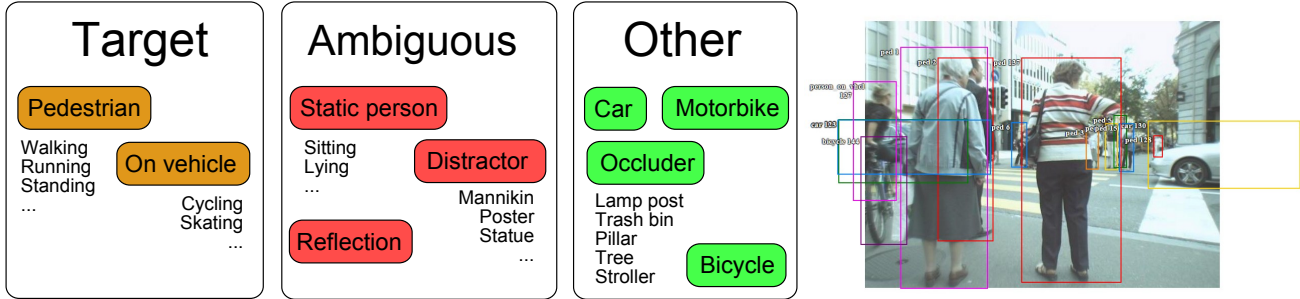Table 7. Overview of the types of annotations currently found in the *MOT16* benchmark.

Figure 9. Left: An overview of annotated classes. The classes in orange will be the central ones to evaluate on. The red classes include ambiguous cases such that both recovering and missing them will not be penalized in the evaluation. The classes in green are annotated for training purposes and for computing the occlusion level of all pedestrians. Right: An exemplar of an annotated frame. Note how partially cropped objects are also marked outside of the frame. Also note that the bounding box encloses the entire person but not, *e.g.*, the white bag of Pedestrian 1 (bottom left).

ject), then several boxes may be used to better approximate the extent of that object.

Persons on vehicles will only be annotated separately from the vehicle if clearly visible. For example, children inside strollers or people inside cars will not be annotated, while motorcyclists or bikers will be.

Start and end of trajectories

The box (track) appears as soon as the person's location and extent can be determined precisely. This is typically the case when $\approx 10\%$ of the person becomes visible. Similarly,

the track ends when it is no longer possible to pinpoint the exact location. In other words the annotation starts as early and ends as late as possible such that the accuracy is not forfeited. The box coordinates may exceed the visible area. Should a person leave the field of view and appear at a later point, they will be assigned a new ID.

Minimal size

Although the evaluation will only take into account pedestrians that have a minimum height in pixels, annotations will contain all objects of all sizes as long as they are

14

	Instruction
What?	Targets: All upright people including + walking, standing, running pedestrians + cyclists, skaters Distractors: Static people or representations + people not in upright position (sitting, lying down) + reflections, drawings, or photographs of people + human-like objects like dolls, mannequins Others: Moving vehicles and occluders + Cars, bikes, motorbikes + Pillars, trees, buildings
When?	Start as early as possible. End as late as possible. Keep ID as long as the person is inside the field of view and its path can be determined unambiguously.
How?	The bounding box should contain all pixels belonging to that person and at the same time be as tight as possible.
Occlusions	Always annotate during occlusions if the position can be determined unambiguously. If the occlusion is very long and it is not possible to determine the path of the object using simple reasoning (*e.g.*, constant velocity assumption), the object will be assigned a new ID once it reappears.

Table 8. Instructions obeyed during annotations.

distinguishable by the annotator. In other words, *all* targets independent of their size on the image shall be annotated.

Occlusions

There is no need to explicitly annotate the level of occlusion. This value will be computed automatically using the ground plane assumption and the annotations. Each target is fully annotated through occlusions as long as its extent and location can be determined accurately enough. If a target becomes completely occluded in the middle of the sequence and does not become visible later, the track should be terminated (marked as 'outside of view'). If a target reappears after a prolonged period such that its location is ambiguous during the occlusion, it will reappear with a new ID.

Sanity check

Upon annotating all sequences, a "sanity check" was carried out to ensure that no relevant entities were missed. To that end, we ran a pedestrian detector on all videos and added all high-confidence detections that corresponded to

either humans or distractors to the annotation list.

Runtime Analysis

Different methods require a varying degree of computational resources to perform the task of multi-target tracking. In our current setup, we only consider the final tracking output in form of labeled bounding boxes and therefore cannot directly measure the efficiency of a particular method. Moreover, the efficiency is extremely hard to compare because some methods may require large amounts of memory, others prefer a multi-core system, while others still can be easily executed on a GPU. For our purpose, we ask each benchmark participant to provide the number of seconds required to produce the results on the entire dataset. The resulting numbers are therefore only indicative of each approach and are not immediately comparable to one another.

Fig. 10 shows the relationship between each submission's performance measured by MOTA and its efficiency in terms of frames per second, averaged over the entire dataset. There are two observations worth pointing out. First, with very few exceptions, the majority of methods is still below real-time performance, which is assumed at 25 Hz. Second, the average processing is slower for the *MOT16* dataset, which is most likely due to its higher average density of 26 persons per frame, compared to only 9 in *MOT15*.

Error Analysis

Here, we provide additional figures for Sections 4 and 5 of the main paper.

False negatives *vs*. false positives

One of the rather surprising findings of our work is the fact that most tracking methods aim to reduce false alarms produced by the person detector and hardly focus on increasing recall, *i.e.* reducing the number of false negatives. This is evidenced by Figures 2 and 4 in the paper. Fig. 11 visualizes the detector coverage as described in Sec. 4.2. For each detection threshold, we computed the percentage of trajectories covered by detections. A ground truth track is considered 'covered' if there is at least one detection with an overlap over 50%. Taking all detections with the confidence value above -0.5 into account, as many as 18% of the trajectories are not covered by any detection. What is even more surprising is that if we drop the 10% detections with the lowest confidence and only consider those with the confidence value above 0.2, the number of completely uncovered tracks jumps up to 55%. Therefore we can conclude that reducing the number of false alarms by choosing a more conservative threshold will significantly hurt the tracking performance.
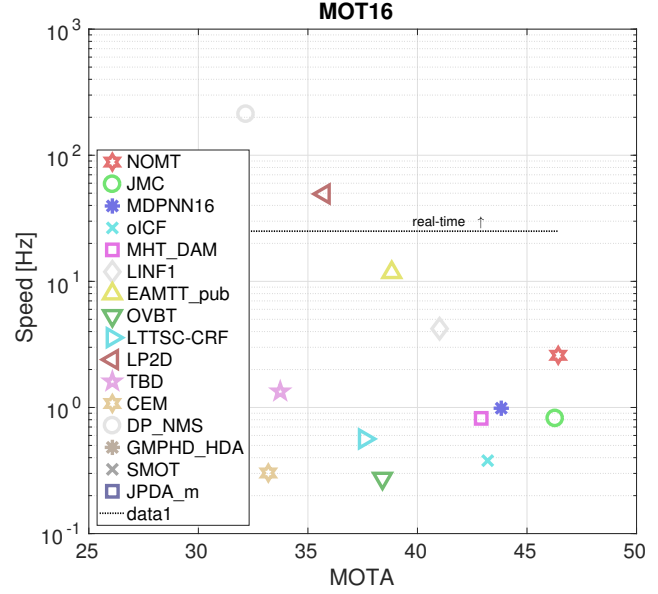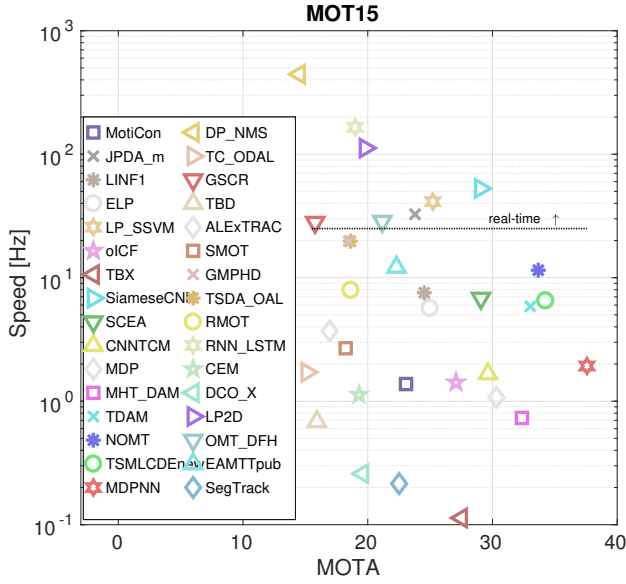
15

Figure 10. Tracker performance measured by MOTA *vs.* processing efficiency in frames per second for *MOT15* (left) and *MOT16* on a log-scale. The latter is only indicative of the true value and has not been measured by the benchmark organizers. See text for details.

Fine-grained Analysis

We include additional plots depicting the fine-grained analysis in Fig. 12.

State of the Art in Multi-Object Tracking

Detailed tracking results for both *MOT15* and *MOT16* datasets are presented in Tables 9 and 2, respectively. In addition to the standard metrics described in Section 6 of the paper, we also list the false alarm per frame (FAF) rate, and the number of track fragmentations (Frag). A track frag-

mentation is counted when a ground truth trajectory is lost for any number of frames after successfully being tracked and before tracking is resumed. For MOTA, we also provide the standard deviation across all sequences as an indicator for the stability of the tracker. Next to the raw numbers of ID switches and fragmentations, we also list the relative errors in parentheses, which are computed as IDSw/Recall and Frag/Recall, respectively. This is important to take into account the number of correctly tracked targets.
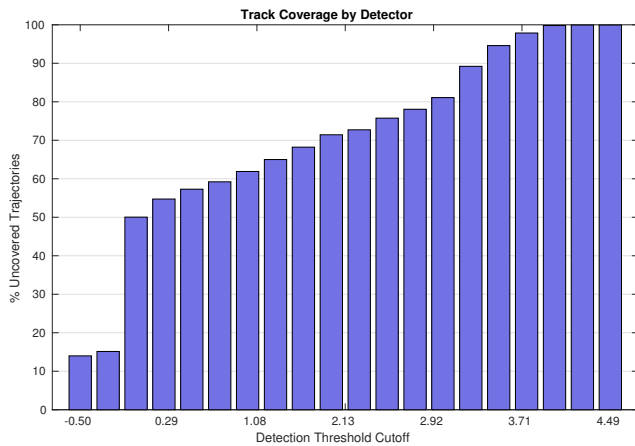

Figure 11. Percentage of covered trajectories (with at least one detection with 50% overlap or more) when varying the detection threshold cutoff. Note that if we drop the 10% detections with lowest confidence, more than 50% of the trajectories are not covered by any detection.
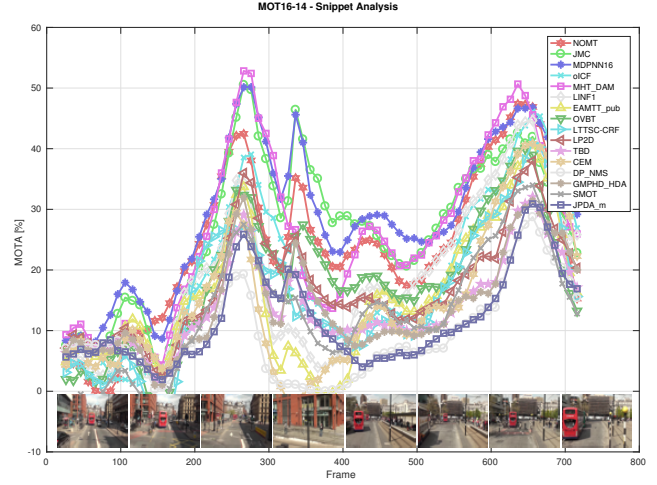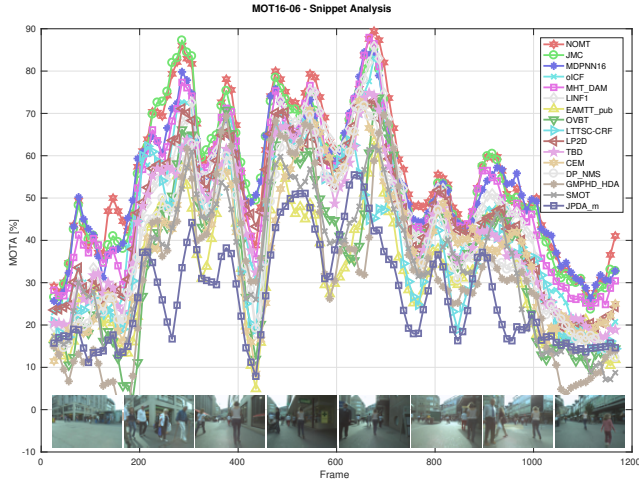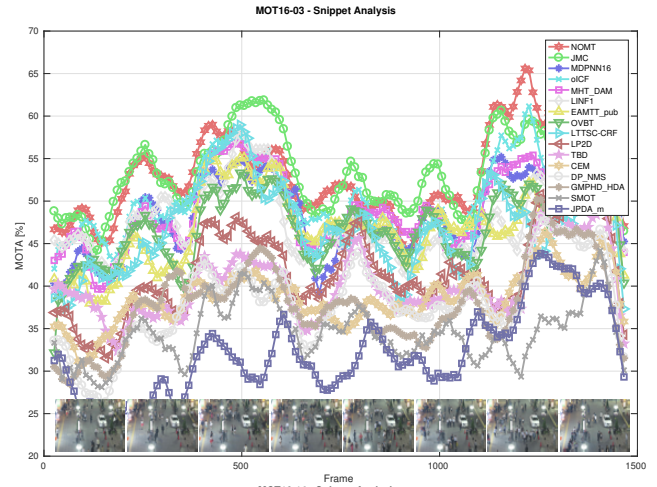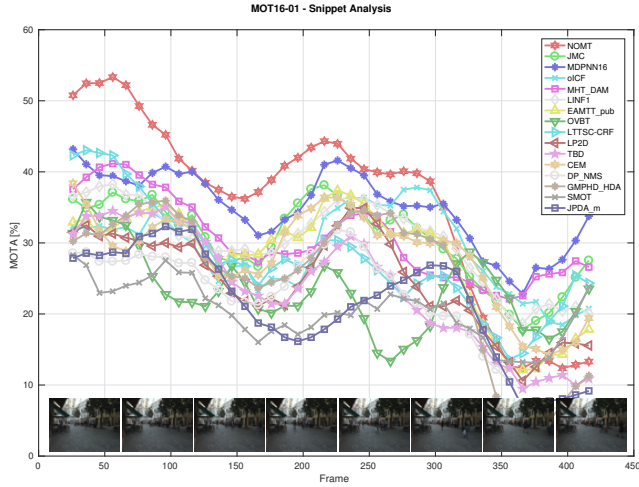
16

Figure 12. Additional examples of fine-grained analysis on four sequences. Each line represents the *local* performance of a tracker, measured by MOTA within a 50-frame (roughly 3 seconds long) segment. Note the extreme within-sequence variation, in particular for the two moving-camera sequences on the bottom.

Table 9. The *MOT15* leaderboard. Performance of several trackers according to different metrics.

Method	MOTA	MOTP	FAF	MT	ML	FP	FN	IDsw	Frag
MDPNN [50]	36.6 ± 12.1	71.4	1.1	13.3	36.5	6419	31811	700 (14.5)	1458 (30.2)
TSMLCDEnew [59]	34.3±13.1	71.7	1.4	14.0	39.4	7869	31908	618 (12.9)	959 (20.0)
NOMT [8]	33.7± 16.2	71.9	1.3	12.2	44.0	7762	32547	442 (9.4)	823 (17.5)
TDAM [44]	33.0±9.8	72.8	1.7	13.3	39.1	10064	30617	464 (9.2)	1506 (30.0)
MHT_DAM [29]	32.4±15.6	71.8	1.6	16.0	43.8	9064	32060	435 (9.1)	826 (17.3)
MDP [61]	30.3± 14.6	71.3	1.7	13.0	38.4	9717	32422	680 (14.4)	1500 (31.8)
CNNTCM [58]	29.6±13.9	71.8	1.3	11.2	44.0	7786	34733	712 (16.4)	943 (21.7)
SCEA [62]	29.1±12.2	71.1	1	8.9	47.3	6060	36912	604 (15.1)	1182 (29.6)
SiameseCNN [34]	29.0±15.1	71.2	0.9	8.5	48.4	5160	37798	639 (16.6)	1316 (34.2)
TbX [24]	27.5±13.3	70.6	1.4	10.4	45.8	7968	35810	759 (18.2)	1528 (36.6)
olCF [28]	27.1±14.9	70	1.3	6.4	48.7	7594	36757	454 (11.3)	1660 (41.3)
LP_SSVM [60]	25.2±13.7	71.7	1.4	5.8	53.0	8369	36932	646 (16.2)	849 (21.3)
ELP [38]	25.0±10.8	71.2	1.3	7.5	43.8	7345	37344	1396 (35.6)	1804 (46.0)
LINF1 [18]	24.5±15.4	71.3	1	5.5	64.6	5864	40207	298 (8.6)	744 (21.5)
JPDA_m [48]	23.8±15.1	68.2	1.1	5.0	58.1	6373	40084	365 (10.5)	869 (25.0)
MotiCon [35]	23.1±16.4	70.9	1.8	4.7	52.0	10404	35844	1018 (24.4)	1061 (25.5)
SegTrack [39]	22.5±15.2	71.7	1.4	5.8	63.9	7890	39020	697 (19.1)	737 (20.2)
EAMTTpub [51]	22.3±14.2	70.8	1.4	5.4	52.7	7924	38982	833 (22.8)	1485 (40.6)
OMT_DFH [26]	21.2±17.2	69.9	2.3	7.1	46.5	13218	34657	563 (12.9)	1255 (28.8)
LP2D [36]	19.8±14.2	71.2	2	6.7	41.2	11580	36045	1649 (39.9)	1712 (41.4)
DCO_X [43]	19.6±14.1	71.4	1.8	5.1	54.9	10652	38232	521 (13.8)	819 (21.7)
CEM [41]	19.3±17.5	70.7	2.5	8.5	46.5	14180	34591	813 (18.6)	1023 (23.4)
RMM_LSTM [40]	19.0±15.2	71	2	5.5	45.6	11578	36706	1490 (37.0)	2081 (51.7)
RMOT [63]	18.6±17.5	69.6	2.2	5.3	53.3	12473	36835	684 (17.1)	1282 (32.0)
TSDA_OAL [30]	18.6±17.6	69.7	2.8	9.4	42.3	16350	32853	806 (17.3)	1544 (33.2)
GMPHD_15 [55]	18.5±12.7	70.9	1.4	3.9	55.3	7864	41766	459 (14.3)	1,266 (39.5)
SMOT [11]	18.2±10.3	71.2	1.5	2.8	54.8	8780	40310	1148 (33.4)	2132 (62.0)
ALExTRAC [6]	17.0±12.1	71.2	1.6	3.9	52.4	9233	39933	1859 (53.1)	1872 (53.5)
TBD [21]	15.9±17.6	70.9	2.6	6.4	47.9	14943	34777	1939 (44.7)	1963 (45.2)
GSCR [17]	15.8±10.5	69.4	1.3	1.8	61.0	7597	43633	514 (17.7)	1010 (34.8)
TC_ODAL [3]	15.1±15.0	70.5	2.2	3.2	55.8	12970	38538	637 (17.1)	1716 (46.0)
DP_NMS [46]	14.5±13.9	70.8	2.3	6.0	40.8	13171	34814	4537 (104.7)	3090 (71.3)