MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking

Author(s):
Dendorfer, Patrick; Osep, Aljosa; Milan, Anton; Schindler, Konrad; Cremers, Daniel; Reid, Ian; Roth, Stefan; Leal-Taixé, Laura

MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking

Patrick Dendorfer[1] · Aljoša Ošep[1] · Anton Milan[2] · Konrad Schindler[3] · Daniel Cremers[1] · Ian Reid[4] · Stefan Roth[5] · Laura Leal-Taixé[1]

Abstract
Standardized benchmarks have been crucial in pushing the performance of computer vision algorithms, especially since the advent of deep learning. Although leaderboards should not be over-claimed, they often provide the most objective measure of performance and are therefore important guides for research. We present *MOTChallenge*, a benchmark for single-camera Multiple Object Tracking (MOT) launched in late 2014, to collect existing and new data and create a framework for the standardized evaluation of multiple object tracking methods. The benchmark is focused on multiple people tracking, since pedestrians are by far the most studied object in the tracking community, with applications ranging from robot navigation to self-driving cars. This paper collects the first three releases of the benchmark: (i) *MOT15*, along with numerous state-of-the-art results that were submitted in the last years, (ii) *MOT16*, which contains new challenging videos, and (iii) *MOT17*, that extends *MOT16* sequences with more precise labels and evaluates tracking performance on three different object detectors. The second and third release not only offers a significant increase in the number of labeled boxes, but also provide labels for multiple object classes beside pedestrians, as well as the level of visibility for every single object of interest. We finally provide a categorization of state-of-the-art trackers and a broad error analysis. This will help newcomers understand the related work and research trends in the MOT community, and hopefully shed some light into potential future research directions.

Keywords Multi-object-tracking · Evaluation · MOTChallenge · Computer vision · MOTA

Communicated by Daniel Scharstein.

Anton Milan: Work done prior to joining Amazon.

✉ Patrick Dendorfer
patrick.dendorfer@tum.de

Aljoša Ošep
aljosa.osep@tum.de

Anton Milan
antmila@amazon.com

Konrad Schindler
schindler@ethz.ch

Daniel Cremers
cremers@tum.de

Ian Reid
ian.reid@adelaide.edu.au

Stefan Roth
stefan.roth@vision.tu-darmstadt.de

Laura Leal-Taixé
leal.taixe@tum.de

1 Introduction

Evaluating and comparing single-camera multi-target tracking methods is not trivial for numerous reasons (Milan et al. 2013). Firstly, unlike for other tasks, such as image denoising, 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. Secondly, many different evaluation metrics with free parameters and ambiguous definitions often lead to conflicting quantitative results across

[1] Technical University Munich, Munich, Germany

[2] Amazon Research, Tübingen, Germany

[3] ETH Zürich, Zurich, Switzerland

[4] The University of Adelaide, Adelaide, Australia

[5] Technical University of Darmstadt, Darmstadt, Germany

the literature. Finally, the lack of pre-defined test and training data makes it difficult to compare different methods fairly.

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 for each sequence, reaching over 90% accuracy in well-known sequences like PETS (Ferryman and Ellis 2010). Nonetheless, the real challenge for a tracking system is to be able to perform well on a variety of sequences with different level of crowdedness, camera motion, illumination, etc., without overfitting the set of parameters to a specific video sequence.

To address this issue, we released the *MOTChallenge* benchmark in 2014, which consisted of three main components: (1) a (re-)collection of publicly available and new datasets, (2) a centralized evaluation method, and (3) an infrastructure that allows for crowdsourcing of new data, new evaluation methods and even new annotations. The first release of the dataset named *MOT15* consists of 11 sequences for training and 11 for testing, with a total of 11286 frames or 996 seconds of video. 3D information was also provided for 4 of those sequences. Pre-computed object detections, annotations (only for the training sequences), and a common evaluation method for all datasets were provided to all participants, which allowed for all results to be compared fairly.

Since October 2014, over 1,000 methods have been publicly tested on the *MOTChallenge* benchmark, and over 1833 users have registered, see Fig. 1. In particular, 760 methods have been tested on *MOT15*, 1,017 on *MOT16* and 692 on *MOT17*; 132, 213 and 190 (respectively) were published on the public leaderboard. This established *MOTChallenge* as the first standardized large-scale tracking benchmark for single-camera multiple people tracking.

Despite its success, the first tracking benchmark, *MOT15*, was lacking in a few aspects:

- The annotation protocol was not consistent across all sequences since some of the ground truth was collected from various online sources;
- the distribution of crowd density was not balanced for training and test sequences;
- some of the sequences were well-known (e.g., PETS09-S2L1) and methods were overfitted to them, which made them not ideal for testing purposes;
- the provided public detections did not show good performance on the benchmark, which made some participants switch to other pedestrian detectors.

To resolve the aforementioned shortcomings, we introduced the second benchmark, *MOT16*. It consists of a set of 14 sequences with crowded scenarios, recorded from different viewpoints, with/without camera motion, and it covers a diverse set of weather and illumination conditions. Most importantly, the annotations for *all* sequences were carried out by qualified researchers from scratch following a strict protocol and finally double-checked to ensure a high annotation accuracy. In addition to pedestrians, we also annotated classes such as vehicles, sitting people, and occluding objects. With this fine-grained level of annotation, it was possible to accurately compute the degree of occlusion and cropping of all bounding boxes, which was also provided with the benchmark.

For the third release, *MOT17*, we (1) further improved the annotation consistency over the sequences[1] and (2) proposed a new evaluation protocol with public detections. In *MOT17*, we provided 3 sets of public detections, obtained using three different object detectors. Participants were required to evaluate their trackers using all three detections sets, and results were then averaged to obtain the final score. The main idea behind this new protocol was to establish the robustness of the trackers when fed with detections of different quality. Besides, we released a separate subset for evaluating object detectors, *MOT17Det*.

In this work, we categorize and analyze 73 published trackers that have been evaluated on *MOT15*, 74 trackers on *MOT16*, and 57 on *MOT17*.[2] Having results on such a large number of sequences allows us to perform a thorough analysis of trends in tracking, currently best-performing methods, and special failure cases. We aim to shed some light on potential research directions for the near future in order to further improve tracking performance.

In summary, this paper has two main goals:

- To present the *MOTChallenge* benchmark for a fair evaluation of multi-target tracking methods, along with its first releases: *MOT15*, *MOT16*, and *MOT17*;
- to analyze the performance of 73 state-of-the-art trackers on *MOT15*, 74 trackers on *MOT16*, and 57 on *MOT17* to analyze trends in MOT over the years. We analyze the main weaknesses of current trackers and discuss promising research directions for the community to advance the field of multi-target tracking.

The benchmark with all datasets, ground truth, detections, submitted results, current ranking and submission guidelines can be found at:

http://www.motchallenge.net/.

---

[1] We thank the numerous contributors and users of MOTChallenge that pointed us to issues with annotations.

[2] In this paper, we only consider published trackers that were on the leaderboard on April 17th, 2020, and used the provided set of public detections. For this analysis, we focused on peer-reviewed methods, i.e., published at a conference or a journal, and excluded entries for which we could not find corresponding publications due to lack of information provided by the authors.

2 Related work

Benchmarks and challenges In the recent past, the computer vision community has developed centralized benchmarks for numerous tasks including object detection (Everingham et al. 2015), pedestrian detection (Dollár et al. 2009), 3D reconstruction (Seitz et al. 2006), optical flow (Baker et al. 2011; Geiger et al. 2012), visual odometry (Geiger et al. 2012), single-object short-term tracking (Kristan et al. 2014), and stereo estimation (Geiger et al. 2012; Scharstein and Szeliski 2002). Despite potential pitfalls of such benchmarks (Torralba and Efros 2011), they have proven to be extremely helpful to advance the state of the art in the respective area.

For single-camera 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 (Ferryman and Ellis 2010) addressing primarily surveillance applications. The 2009 version consists of 3 subsets S: S1 targeting person count and density estimation, S2 targeting people tracking, and S3 targeting flow analysis and event recognition. The simplest sequence for tracking (S2L1) consists of a scene with few pedestrians, and for that sequence, state-of-the-art methods perform extremely well with accuracies of over 90% given a good set of initial detections (Henriques et al. 2011; Milan et al. 2014; Zamir et al. 2012). Therefore, methods started to focus on tracking objects in the most challenging sequence, i.e., with the highest crowd density, but hardly ever on the complete dataset. Even for this widely used benchmark, we observe that tracking results are commonly obtained inconsistently, involving 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. Hence, the questions that arise are: (i) are these sequences already too easy for current tracking methods?, (ii) do methods simply overfit?, and (iii) are existing methods poorly evaluated?

The PETS team organizes a workshop approximately once a year to which researchers can submit their results, and methods are evaluated under the same conditions. Although this is indeed a fair comparison, the fact that submissions are evaluated only once a year means that the use of this benchmark for high impact conferences like ICCV or CVPR remains challenging. Furthermore, the sequences tend to be focused only on surveillance scenarios and lately on specific tasks such as vessel tracking. Surveillance videos have a low frame rate, fixed camera viewpoint, and low pedestrian density. The ambition of *MOTChallenge* is to tackle more general scenarios including varying viewpoints, illumination conditions, different frame rates, and levels of crowdedness.

A well-established and useful way of organizing datasets is through standardized challenges. These are usually in the form of web servers that host the data and through which results are uploaded by the users. Results are then evaluated in a centralized way by the server and afterward presented online to the public, making a comparison with any other method immediately possible.

There are several datasets organized in this fashion: the Labeled Faces in the Wild (Huang et al. 2007) for unconstrained face recognition, the PASCAL VOC (Everingham et al. 2015) for object detection and the ImageNet large scale visual recognition challenge (Russakovsky et al. 2015).

The KITTI benchmark (Geiger et al. 2012) was introduced for challenges in autonomous driving, which includes 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, but the camera position is located at a fixed height for all sequences.

Another work that is worth mentioning is Alahi et al. (2014), in which the authors collected a large amount of data containing 42 million pedestrian trajectories. Since annotation of such a large collection of data is infeasible, they use a denser set of cameras to create the "ground-truth" trajectories. Though we do not aim at collecting such a large amount of data, the goal of our benchmark is somewhat similar: to push research in tracking forward by generalizing the test data to a larger set that is highly variable and hard to overfit.

DETRAC (Wen et al. 2020) is a benchmark for vehicle tracking, following a similar submission system to the one we proposed with *MOTChallenge*. This benchmark consists of a total of 100 sequences, 60% of which are used for training. Sequences are recorded from a high viewpoint (surveillance scenarios) with the goal of vehicle tracking.

Evaluation A critical question with any dataset is how to measure the performance of the algorithms. In the case of multiple object tracking, the CLEAR-MOT metrics (Stiefelhagen et al. 2006) have emerged as the standard measures. By measuring the intersection over union of bounding boxes and matching those from ground-truth annotations and results, measures of accuracy and precision can be computed. Precision measures how well the persons are localized, while accuracy evaluates how many distinct errors such as missed targets, ghost trajectories, or identity switches are made.

Alternatively, trajectory-based measures by Wu and Nevatia (2006) evaluate how many trajectories were mostly tracked, mostly lost, and partially tracked, relative to the track lengths. These are mainly used to assess track coverage. The IDF1 metric (Ristani et al. 2016) was introduced for MOT evaluation in a multi-camera setting. Since then it has been adopted for evaluation in the standard single-camera setting in our benchmark. Contrary to MOTA, the ground truth to predictions mapping is established at the level of entire tracks instead of on frame by frame level, and therefore, measures long-term tracking quality. In Sect. 7 we report IDF1 perfor-
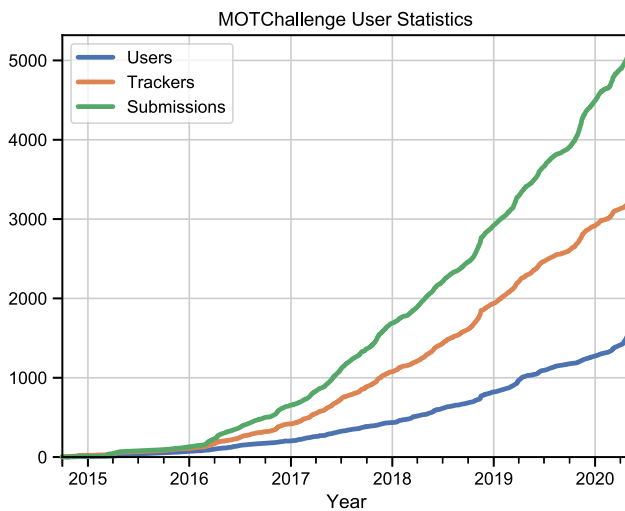
Fig. 1 Evolution of *MOTChallenge* submissions, number of users registered and trackers created

mance in conjunction with MOTA. A detailed discussion on the measures can be found in Sect. 6.

A key parameter in both families of metrics is the intersection over union threshold which determines whether a predicted bounding box was matched to an annotation. It is fairly common to observe methods compared under different thresholds, varying from 25 to 50%. There are often many other variables and implementation details that differ between evaluation scripts, which may affect results significantly. Furthermore, the evaluation script is not the only factor. Recently, a thorough study (Mathias et al. 2014) on face detection benchmarks showed that annotation policies vary greatly among datasets. For example, bounding boxes can be defined tightly around the object, or more loosely to account for pose variations. The size of the bounding box can greatly affect results since the intersection over union depends directly on it.

Standardized benchmarks are preferable for comparing methods in a fair and principled way. Using the same ground-truth data and evaluation methodology is the only way to guarantee that the only part being evaluated is the tracking method that delivers the results. This is the main goal of the *MOTChallenge* benchmark.

3 History of MOTChallenge

The first benchmark was released in October 2014 and it consists of 11 sequences for training and 11 for testing, where the testing sequences have not been available publicly. We also provided a set of detections and evaluation scripts. Since its release, 692 tracking results were submitted to the benchmark, which has quickly become the standard for evaluating multiple pedestrian tracking methods in high impact confer-

ences such as ICCV, CVPR, and ECCV. Together with the release of the new data, we organized the 1st Workshop on Benchmarking Multi-Target Tracking (BMTT) in conjunction with the IEEE Winter Conference on Applications of Computer Vision (WACV) in 2015.[3]

After the success of the first release of sequences, we created a 2016 edition, with 14 longer and more crowded sequences and a more accurate annotation policy which we describe in this manuscript (Sect. C.1). For the release of *MOT16*, we organized the second workshop[4] in conjunction with the European Conference in Computer Vision (ECCV) in 2016.

For the third release of our dataset, *MOT17*, we improved the annotation consistency over the *MOT16* sequences and provided three public sets of detections, on which trackers need to be evaluated. For this release, we organized a Joint Workshop on Tracking and Surveillance in conjunction with the Performance Evaluation of Tracking and Surveillance (PETS) (Ferryman and Ellis 2010; Ferryman and Shahrokni 2009) workshop and the Conference on Vision and Pattern Recognition (CVPR) in 2017.[5]

In this paper, we focus on the *MOT15*, *MOT16*, and *MOT17* benchmarks because numerous methods have already submitted their results to these challenges for several years that allow us to analyze these methods and to draw conclusions about research trends in multi-object tracking.

Nonetheless, work continues on the benchmark, with frequent releases of new challenges and datasets. The latest pedestrian tracking dataset was first presented at the 4th *MOTChallenge* workshop[6] (CVPR 2019), an ambitious tracking challenge with eight new sequences (Dendorfer et al. 2019). With the feedback of the workshop the sequences were revised and re-published as the *MOT20* (Dendorfer et al. 2020) benchmark. This challenge focuses on very crowded scenes, where the object density can reach up to 246 pedestrians per frame. The diverse sequences show indoor and outdoor scenes, filmed either during day or night. With more than $2M$ bounding boxes and 3833 tracks, *MOT20* constitutes a new level of complexity and challenges the performance of tracking methods in very dense scenarios. At the time of this article, only 11 submissions for *MOT20* had been received, hence a discussion of the results is not yet significant nor informative, and is left for future work.

The future vision of *MOTChallenge* is to establish it as a general platform for benchmarking multi-object tracking, expanding beyond pedestrian tracking. To this end, we recently added a public benchmark for multi-camera 3D zebrafish tracking (Pedersen et al. 2020), and a benchmark
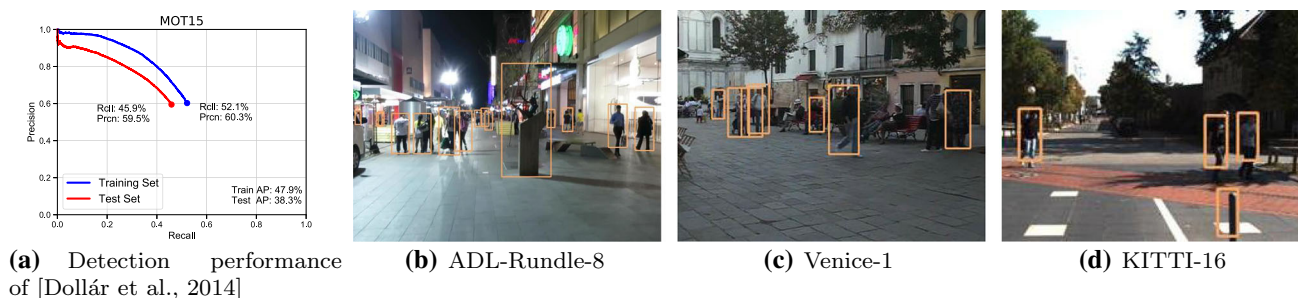
---

3 https://motchallenge.net/workshops/bmtt2015/.

4 https://motchallenge.net/workshops/bmtt2016/.

5 https://motchallenge.net/workshops/bmtt-pets2017/.

6 https://motchallenge.net/workshops/bmtt2019/.

(a) Detection performance of [Dollár et al., 2014]   **(b)** ADL-Rundle-8   **(c)** Venice-1   **(d)** KITTI-16

Fig. 2 **a** The performance of the provided detection bounding boxes evaluated on the training (blue) and the test (red) set. The circle indicates the operating point (i.e., the input detection set) for the trackers. **b–d** Exemplar detection results

for the large-scale Tracking any Object (TAO) dataset (Dave et al. 2020). This dataset consists of 2907 videos, covering 833 classes by 17,287 tracks.

In Fig. 1, we plot the evolution of the number of users, submissions, and trackers created since *MOTChallenge* was released to the public in 2014. Since our 2nd workshop was announced at ECCV, we have experienced steady growth in the number of users as well as submissions.

4 MOT15 Release

One of the key aspects of any benchmark is data collection. The goal of *MOTChallenge* is not only to compile yet another dataset with completely new data but rather to: (1) create a common framework to test tracking methods on, and (2) 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. In Table 5 of the Appendix we show an overview of the sequences included in the benchmark.

4.1 Sequences

We have compiled a total of 22 sequences that combine different videos from several sources (Andriluka et al. 2010; Benfold and Reid 2011; Ess et al. 2008; Ferryman and Ellis 2010; Geiger et al. 2012) and new data collected from us. We use half of the data for training and a half for testing, and the annotations of the testing sequences are not released to the public to avoid (over)fitting of methods to specific sequences. Note, the test data contains over 10 min of footage and 61,440 annotated bounding boxes, therefore, it is hard for researchers to over-tune their algorithms on such a large amount of data. This is one of the major strengths of the benchmark.

We collected 6 new challenging sequences, 4 filmed from a static camera and 2 from a moving camera held at pedestrian's height. Three sequences are particularly challenging: a night sequence filmed 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. A smaller subset of the benchmark including only these six new sequences were presented at the 1st Workshop on Benchmarking Multi-Target Tracking,[7] where the top-performing method reached MOTA (tracking accuracy) of only 12.7%. This confirms the difficulty of the new sequences.[8]

4.2 Detections

To detect pedestrians in all images of the *MOT15* edition, we use the object detector of Dollár et al. (2014), which is based on aggregated channel features (ACF). We rely on the default parameters and the pedestrian model trained on the INRIA dataset (Dalal and Triggs 2005), rescaled with a factor of 0.6 to enable the detection of smaller pedestrians. The detector performance along with three sample frames is depicted in Fig. 2, for both the training and the test set of the benchmark. Recall does not reach 100% because of the non-maximum suppression applied.

We cannot (nor necessarily want to) prevent anyone from using a different set of detections. However, we require that this is noted as part of the tracker's description and is also displayed in the rating table.

4.3 Weaknesses of *MOT15*

By the end of 2015, it was clear that a new release was due for the *MOTChallenge* benchmark. The main weaknesses of *MOT15* were the following:

– *Annotations* we collected annotations online for the existing sequences, while we manually annotated the new sequences. Some of the collected annotations were not

---

[7] https://motchallenge.net/workshops/bmtt2015/.

[8] The challenge results are available at http://motchallenge.net/results/WACV_2015_Challenge/.
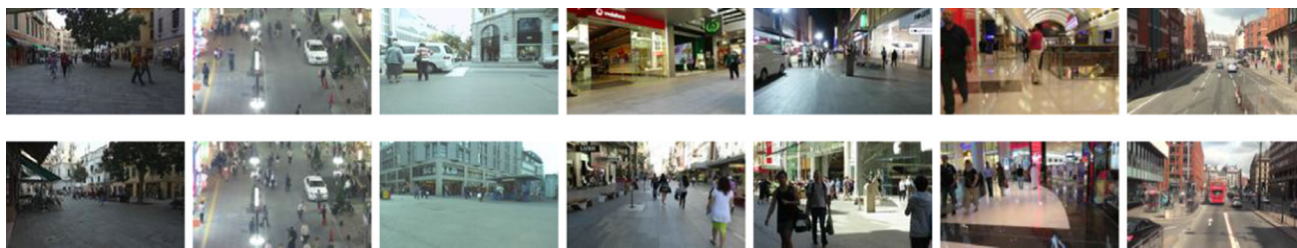
Fig. 3 An overview of the *MOT16/MOT17* dataset. Top: Training sequences. Bottom: test sequences (Color figure online)

(a) DPM v5 **(b)** DPM v5 **(c)** Faster-RCNN **(d)** SDP
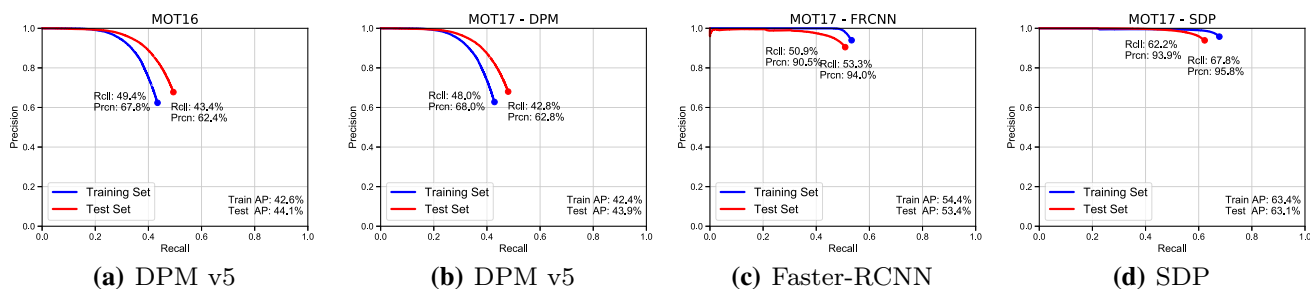
Fig. 4 The performance of three popular pedestrian detectors evaluated on the training (blue) and the test (red) set. The circle indicates the operating point (i.e. the input detection set) for the trackers of *MOT16* and *MOT17* (Color figure online)

accurate enough, especially in scenes with moving cameras.

– *Difficulty* generally, we wanted to include some well-known sequences, e.g., PETS2009, in the *MOT15* benchmark. However, these sequences have turned out to be too simple for state-of-the-art trackers why we concluded to create a new and more challenging benchmark.

To overcome these weaknesses, we created *MOT16*, a collection of all-new challenging sequences (including our new sequences from *MOT15*) and creating annotations following a more strict protocol (see Sect. C.1 of the Appendix).

5 MOT16 and MOT17 Releases

Our ambition for the release of *MOT16* was to compile a benchmark with new and more challenging sequences compared to *MOT15*. Figure 3 presents an overview of the benchmark training and test sequences (detailed information about the sequences is presented in Table 9 in the Appendix).

MOT17 consists of the same sequences as *MOT16*, but contains two important changes: (i) the annotations are further improved, i.e., increasing the accuracy of the bounding boxes, adding missed pedestrians, annotating additional occluders, following the comments received by many anonymous benchmark users, as well as the second round of sanity checks, (ii) the evaluation system significantly differs from *MOT17*, including the evaluation of tracking methods using

three different detectors in order to show the robustness to varying levels of noisy detections.

5.1 MOT16 Sequences

We compiled a total of 14 sequences, of which we use half for training and a half for testing. The annotations of the testing sequences are not publicly available. The sequences can be classified according to moving/static camera, viewpoint, and illumination conditions (Fig. 11 in Appendix). The new data contains almost 3 times more bounding boxes for training and testing than *MOT15*. Most sequences are filmed in high resolution, and the mean crowd density is 3 times higher when compared to the first benchmark release. Hence, the new sequences present a more challenging benchmark than *MOT15* for the tracking community.

5.2 Detections

We evaluate several state-of-the-art detectors on our benchmark, and summarize the main findings in Fig. 4. To evaluate the performance of the detectors for the task of tracking, we evaluate them using all bounding boxes considered for the tracking evaluation, including partially visible or occluded objects. Consequently, the recall and average precision (AP) is lower than the results obtained by evaluating solely on visible objects, as we do for the detection challenge.

MOT16 Detections We first train the deformable part-based model (DPM) v5 (Felzenszwalb and Huttenlocher 2006) and find that it outperforms other detectors such as Fast-

RNN (Girshick 2015) and ACF (Dollár et al. 2014) for the task of detecting persons on *MOT16*. Hence, for that benchmark, we provide DPM detections as public detections.

MOT17 Detections For the new *MOT17* release, we use Faster-RCNN (Ren et al. 2015) and a detector with scale-dependent pooling (SDP) (Yang et al. 2016), both of which outperform the previous DPM method. After a discussion held in one of the *MOTChallenge* workshops, we agreed to provide all three detections as public detections, effectively changing the way *MOTChallenge* evaluates trackers. The motivation is to challenge trackers further to be more general and work with detections of varying quality. These detectors have different characteristics, as can be seen in in Fig. 4. Hence, a tracker that can work with all three inputs is going to be inherently more robust. The evaluation for *MOT17* is, therefore, set to evaluate the output of trackers on all three detection sets, averaging their performance for the final ranking. A detailed breakdown of detection bounding box statistics on individual sequences is provided in Table 10 in the Appendix.

6 Evaluation

MOTChallenge is also a platform for a fair comparison of state-of-the-art tracking methods. By providing authors with standardized ground-truth data, evaluation metrics, scripts, as well as a set of precomputed detections, all methods are compared under the same conditions, thereby isolating the performance of the tracker from other factors. In the past, a large number of metrics for quantitative evaluation of multiple target tracking have been proposed (Bernardin and Stiefelhagen 2008; Li et al. 2009; Schuhmacher et al. 2008; Smith et al. 2005; Stiefelhagen et al. 2006; Wu and Nevatia 2006). Choosing "the right" one is largely application dependent and the quest for a unique, general evaluation measure is still ongoing. On the one hand, it is desirable to summarize the performance into a single number to enable a direct comparison between methods. On the other hand, one might want to provide more informative performance estimates by detailing the types of errors the algorithms make, which precludes a clear ranking.

Following a recent trend (Bae and Yoon 2014; Milan et al. 2014; Wen et al. 2014), we employ three sets of tracking performance measures that have been established in the literature: (i) the frame-to-frame based *CLEAR-MOT* metrics proposed by Stiefelhagen et al. (2006), (ii) track quality measures proposed by Wu and Nevatia (2006), and (iii) trajectory-based IDF1 proposed by Ristani et al. (2016).

These evaluation measures give a complementary view on tracking performance. The main representative of CLEAR-MOT measures, Multi-Object Tracking Accuracy (MOTA), is evaluated based on frame-to-frame matching between track

predictions and ground truth. It explicitly penalizes identity switches between consecutive frames, thus evaluating tracking performance only locally. This measure tends to put more emphasis on object detection performance compared to temporal continuity. In contrast, track quality measures (Wu and Nevatia 2006) and IDF1 Ristani et al. (2016), perform prediction-to-ground-truth matching on a trajectory level and over-emphasize the temporal continuity aspect of the tracking performance. In this section, we first introduce the matching between predicted track and ground-truth annotation before we present the final measures. All evaluation scripts used in our benchmark are publicly available.[9]

6.1 Multiple Object Tracking Accuracy

MOTA summarizes three sources of errors with a single performance measure:

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t}, \qquad (1)$$

where t is the frame index and GT is the number of ground-truth objects. where FN are the false negatives, i.e., the number of ground truth objects that were not detected by the method. FP are the false positives, i.e., the number of objects that were falsely detected by the method but do not exist in the ground-truth. $IDSW$ is the number of identity switches, i.e., how many times a given trajectory changes from one ground-truth object to another. The computation of these values as well as other implementation details of the evaluation tool are detailed in Appendix Sect. D. We report the percentage MOTA $(-\infty, 100]$ in our benchmark. Note, that MOTA can also be negative in cases where the number of errors made by the tracker exceeds the number of all objects in the scene.

Justification We note that MOTA has been criticized in the literature for not having different sources of errors properly balanced. However, to this day, MOTA is still considered to be the most expressive measure for single-camera MOT evaluation. It was widely adopted for ranking methods in more recent tracking benchmarks, such as PoseTrack (Andriluka et al. 2018), KITTI tracking (Geiger et al. 2012), and the newly released Lyft (Kesten et al. 2019), Waymo (Sun et al. 2020), and ArgoVerse (Chang et al. 2019) benchmarks. We adopt MOTA for ranking, however, we recommend taking alternative evaluation measures (Ristani et al. 2016; Wu and Nevatia 2006) into the account when assessing the tracker's performance.

Robustness One incentive behind compiling this benchmark was to reduce dataset bias by keeping the data as diverse as possible. The main motivation is to challenge state-of-the-art

---

[9] http://motchallenge.net/devkit.

approaches and analyze their performance in unconstrained environments and on unseen data. Our experience shows that most methods can be heavily overfitted on one particular dataset, and may not be general enough to handle an entirely different setting without a major change in parameters or even in the model.

6.2 Multiple Object Tracking Precision

The Multiple Object Tracking Precision is the average dissimilarity between all true positives and their corresponding ground-truth targets. For bounding box overlap, this is computed as:

$$\text{MOTP} = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t}, \tag{2}$$

where c_t denotes the number of matches in frame t and $d_{t,i}$ is the bounding box overlap of target i with its assigned ground-truth object in frame t. MOTP thereby gives the average overlap of t_d between all correctly matched hypotheses and their respective objects and ranges between $t_d := 50\%$ and 100%.

It is important to point out that MOTP is a measure of localisation precision, *not* to be confused with the *positive predictive value* or *relevance* in the context of precision / recall curves used, e.g., in object detection.

In practice, it quantifies the localization precision of the detector, and therefore, it provides little information about the actual performance of the tracker.

6.3 Identification Precision, Identification Recall, and F1 Score

CLEAR-MOT evaluation measures provide event-based tracking assessment. In contrast, the IDF1 measure (Ristani et al. 2016) is an identity-based measure that emphasizes the track identity preservation capability over the entire sequence. In this case, the predictions-to-ground-truth mapping is established by solving a bipartite matching problem, connecting pairs with the largest temporal overlap. After the matching is established, we can compute the number of True Positive IDs (IDTP), False Negative IDs (IDFN), and False Positive IDs (IDFP), that generalise the concept of per-frame TPs, FNs and FPs to tracks. Based on these quantities, we can express the Identification Precision (IDP) as:

$$\text{IDP} = \frac{IDTP}{IDTP + IDFP}, \tag{3}$$

and Identification Recall (IDR) as:

$$\text{IDR} = \frac{IDTP}{IDTP + IDFN}. \tag{4}$$

Note that IDP and IDR are the fraction of computed (ground-truth) detections that are correctly identified. IDF1 is then expressed as a ratio of correctly identified detections over the average number of ground-truth and computed detections and balances identification precision and recall through their harmonic mean:

$$\text{IDF1} = \frac{2 \cdot IDTP}{2 \cdot IDTP + IDFP + IDFN}. \tag{5}$$

6.4 Track Quality Measures

The final measures that we report on our benchmark are qualitative, and evaluate the percentage of the ground-truth trajectory that is recovered by a tracking algorithm. Each ground-truth trajectory can be consequently classified as mostly tracked (MT), partially tracked (PT), and mostly lost (ML). As defined in Wu and Nevatia (2006), a target is mostly tracked if it is successfully tracked for at least 80% of its life span, and considered lost in case it is covered for less than 20% of its total length. The remaining tracks are considered to be partially tracked. A higher number of MT and a few ML is desirable. Note, that it is irrelevant for this measure whether the ID remains the same throughout the track. We report MT and ML as a ratio of mostly tracked and mostly lost targets to the total number of ground-truth trajectories.

In certain situations, one might be interested in obtaining long, persistent tracks without trajectory gaps. To that end, the number of track fragmentations (FM) counts how many times a ground-truth trajectory is interrupted (untracked). A fragmentation event happens each time a trajectory changes its status from tracked to untracked and is resumed at a later point. Similarly to the ID switch ratio (c.f. Sect. D.1), we also provide the relative number of fragmentations as FM/Recall.

7 Analysis of State-of-the-Art Trackers

We now present an analysis of recent multi-object tracking methods that submitted to the benchmark. This is divided into two parts: (i) categorization of the methods, where our goal is to help young scientists to navigate the recent MOT literature, and (ii) error and runtime analysis, where we point out methods that have shown good performance on a wide range of scenes. We hope this can eventually lead to new promising research directions.

We consider all valid submissions to all three benchmarks that were published before April 17th, 2020, and used the provided set of public detections. For this analysis, we focus on methods that are peer-reviewed, i.e., published at a conference or a journal. We evaluate a total of 101 (public) trackers; 73 trackers were tested on *MOT15*, 74 on *MOT16* and 57 on

Table 1 The *MOT15* leaderboard

Method	MOTA	IDF1	MOTP	FAR	MT	ML	FP	FN	IDSW	FM	IDSWR	FMR
MPNTrack (Brasó and Leal-Taixé 2020)	51.54	58.61	76.05	1.32	225	187	7620	21,780	375	872	5.81	13.51
Tracktor++v2 (Bergmann et al. 2019)	46.60	47.57	76.36	0.80	131	201	4624	26,896	1290	1702	22.94	30.27
TrctrD15 (Xu et al. 2020)	44.09	45.99	75.26	1.05	124	192	6085	26,917	1347	1868	23.97	33.24
Tracktor++ (Bergmann et al. 2019)	44.06	46.73	75.03	1.12	130	189	6477	26,577	1318	1790	23.23	31.55
KCF (Chu et al. 2019)	38.90	44.54	70.56	1.27	120	227	7321	29,501	720	1440	13.85	27.70
AP_HWDPL_p (Long et al. 2017)	38.49	47.10	72.56	0.69	63	270	4005	33,203	586	1263	12.75	27.48
STRN (Xu et al. 2019)	38.06	46.62	72.06	0.94	83	241	5451	31,571	1033	2665	21.25	54.82
AMIR15 (Sadeghian et al. 2017)	37.57	46.01	71.66	1.37	114	193	7933	29,397	1026	2024	19.67	38.81
JointMC (Keuper et al. 2018)	35.64	45.12	71.90	1.83	167	283	10,580	28,508	457	969	8.53	18.08
RAR15pub (Fang et al. 2018)	35.11	45.40	70.94	1.17	94	305	6771	32,717	381	1523	8.15	32.58
HybridDAT (Yang et al. 2017)	34.97	47.72	72.57	1.46	82	304	8455	31,140	358	1267	7.26	25.69
INARLA (Wu et al. 2019)	34.69	42.06	70.72	1.71	90	216	9855	29,158	1112	2848	21.16	54.20
STAM (Chu et al. 2017)	34.33	48.26	70.55	0.89	82	313	5154	34,848	348	1463	8.04	33.80
QuadMOT (Son et al. 2017)	33.82	40.43	73.42	1.37	93	266	7898	32,061	703	1430	14.70	29.91
NOMT (Choi 2015)	33.67	44.55	71.94	1.34	88	317	7762	32,547	442	823	9.40	17.50
DCCRF (Zhou et al. 2018a)	33.62	39.08	70.91	1.02	75	271	5917	34,002	866	1566	19.39	35.07
TDAM (Yang and Jia 2016)	33.03	46.05	72.78	1.74	96	282	10,064	30,617	464	1506	9.25	30.02
CDA_DDALpb (Bae and Yoon 2018)	32.80	38.79	70.70	0.86	70	304	4983	35,690	614	1583	14.65	37.77
MHT_DAM (Kim et al. 2015)	32.36	45.31	71.83	1.57	115	316	9064	32,060	435	826	9.10	17.27
LFNF (Sheng et al. 2017)	31.64	33.10	72.03	1.03	69	301	5943	35,095	961	1106	22.41	25.79
GMPHD_OGM (Song et al. 2019)	30.72	38.82	71.64	1.13	83	275	6502	35,030	1034	1351	24.05	31.43
PHD_GSDL (Fu et al. 2018)	30.51	38.82	71.20	1.13	55	297	6534	35,284	879	2208	20.65	51.87
MDP (Xiang et al. 2015)	30.31	44.68	71.32	1.68	94	277	9717	32,422	680	1500	14.40	31.76
MCF_PHD (Wojke and Paulus 2016)	29.89	38.18	71.70	1.54	86	317	8892	33,529	656	989	14.44	21.77
CNNTCM (Wang et al. 2016)	29.64	36.82	71.78	1.35	81	317	7786	34,733	712	943	16.38	21.69
RSCNN (Mahgoub et al. 2017)	29.50	36.97	73.07	2.05	93	262	11,866	30,474	976	1176	19.36	23.33
TBSS15 (Zhou et al. 2018b)	29.21	37.23	71.28	1.05	49	316	6068	36,779	649	1508	16.17	37.57
SCEA (Yoon et al. 2016)	29.08	37.15	71.11	1.05	64	341	6060	36,912	604	1182	15.13	29.61
SiameseCNN (Leal-Taixe et al. 2016)	29.04	34.27	71.20	0.89	61	349	5160	37798	639	1316	16.61	34.20

Table 1 continued

Method	MOTA	IDF1	MOTP	FAR	MT	ML	FP	FN	IDSW	FM	IDSWR	FMR
HAM_INTP15 (Yoon et al. 2018a)	28.62	41.45	71.13	1.30	72	317	7485	35,910	460	1038	11.07	24.98
GMMA_intp (Song et al. 2018)	27.32	36.59	70.92	1.36	47	311	7848	35,817	987	1848	23.67	44.31
oICF (Kieritz et al. 2016)	27.08	40.49	69.96	1.31	46	351	7594	36,757	454	1660	11.30	41.32
TO (Manen et al. 2016)	25.66	32.74	72.17	0.83	31	414	4779	40,511	383	600	11.24	17.61
LP_SSVM (Wang and Fowlkes 2016)	25.22	34.05	71.68	1.45	42	382	8369	36,932	646	849	16.19	21.28
HAM_SADF (Yoon et al. 2018a)	25.19	37.80	71.38	1.27	41	420	7330	38,275	357	745	9.47	19.76
ELP (McLaughlin et al. 2015)	24.99	26.21	71.17	1.27	54	316	7345	37,344	1396	1804	35.60	46.00
AdTobKF (Loumponias et al. 2018)	24.82	34.50	70.78	1.07	29	375	6201	39,321	666	1300	18.50	36.11
LINF1 (Fagot-Bouquet et al. 2016)	24.53	34.82	71.33	1.01	40	466	5864	40,207	298	744	8.62	21.53
TENSOR (Shi et al. 2018)	24.32	24.13	71.58	1.15	40	336	6644	38,582	1271	1304	34.16	35.05
TFMOT (Boragule and Jeon 2017)	23.81	32.30	71.35	0.78	35	447	4533	41,873	404	792	12.69	24.87
JPDA_m (Rezatofighi et al. 2015)	23.79	33.77	68.17	1.10	36	419	6373	40,084	365	869	10.50	25.00
MotiCon (Leal-Taixé et al. 2014)	23.07	29.38	70.87	1.80	34	375	10,404	35,844	1018	1061	24.44	25.47
DEEPDA_MOT (Yoon et al. 2019a)	22.53	25.92	70.92	1.27	46	447	7346	39,092	1159	1538	31.86	42.28
SegTrack (Milan et al. 2015)	22.51	31.48	71.65	1.36	42	461	7890	39,020	697	737	19.10	20.20
EAMTTpub (Sanchez-Matilla et al. 2016)	22.30	32.84	70.79	1.37	39	380	7924	38,982	833	1485	22.79	40.63
SAS_MOT15 (Maksai and Fua 2019)	22.16	27.15	71.10	0.97	22	444	5591	41,531	700	1240	21.60	38.27
OMT_DFH (Ju et al. 2017a)	21.16	37.34	69.94	2.29	51	335	13,218	34,657	563	1255	12.92	28.79
MTSTracker (Nguyen Thi Lan Anh et al. 2017)	20.64	31.87	70.32	2.62	65	266	15,161	32,212	1387	2357	29.16	49.55
TC_SIAMESE (Yoon et al. 2018b)	20.22	32.59	71.09	1.06	19	487	6127	42,596	294	825	9.59	26.90
DCO_X (Milan et al. 2016)	19.59	31.45	71.39	1.84	37	396	10,652	38,232	521	819	13.79	21.68
CEM (Milan et al. 2014)	19.30	N/A	70.74	2.45	61	335	14,180	34,591	813	1023	18.60	23.41
RNN_LSTM (Milan et al. 2017)	18.99	17.12	70.97	2.00	40	329	11,578	36,706	1490	2081	37.01	51.69
RMOT (Yoon et al. 2015)	18.63	32.56	69.57	2.16	38	384	12,473	36,835	684	1282	17.08	32.01
TSDA_OAL (Ju et al. 2017b)	18.61	36.07	69.68	2.83	68	305	16,350	32,853	806	1544	17.32	33.18
GMPHD_15 (Song and Jeon 2016)	18.47	28.38	70.90	1.36	28	399	7864	41,766	459	1266	14.33	39.54
SMOT (Dicle et al. 2013)	18.23	0.00	71.23	1.52	20	395	8780	40,310	1148	2132	33.38	61.99
ALExTRAC (Bewley et al. 2016b)	16.95	17.30	71.18	1.60	28	378	9233	39,933	1859	1872	53.11	53.48
TBD (Geiger et al. 2014)	15.92	0.00	70.86	2.58	46	345	14,943	34,777	1939	1963	44.68	45.23
GSCR (Fagot-Bouquet et al. 2015)	15.78	27.90	69.38	1.31	13	440	7597	43,633	514	1010	17.73	34.85
TC_ODAL (Bae and Yoon 2014)	15.13	0.00	70.53	2.24	23	402	12,970	38,538	637	1716	17.09	46.04
DP_NMS (Pirsiavash et al. 2011)	14.52	19.69	70.76	2.28	43	294	13,171	34,814	4537	3090	104.69	71.30

Performance of several trackers according to different metrics

Table 2 The *MOT16* leaderboard

Method	MOTA	IDF1	MOTP	FAR	MT	ML	FP	FN	IDSW	FM	IDSWR	FMR
MPNTrack (Brasó and Leal-Taixé 2020)	58.56	61.69	78.88	0.84	207	258	4949	70,252	354	684	5.76	11.13
Tracktor++v2 (Bergmann et al. 2019)	56.20	54.91	79.20	0.40	157	272	2394	76,844	617	1068	10.66	18.46
TrctrD16 (Xu et al. 2020)	54.83	53.39	77.47	0.50	145	281	2955	78,765	645	1515	11.36	26.67
Tracktor++ (Bergmann et al. 2019)	54.42	52.54	78.22	0.55	144	280	3280	79,149	682	1480	12.05	26.15
NOTA_16 (Chen et al. 2019)	49.83	55.33	74.49	1.22	136	286	7248	83,614	614	1372	11.34	25.34
HCC (Ma et al. 2018b)	49.25	50.67	79.00	0.90	135	303	5333	86,795	391	535	7.46	10.21
eTC (Wang et al. 2019)	49.15	56.11	75.49	1.42	131	306	8400	83,702	606	882	11.20	16.31
KCF16 (Chu et al. 2019)	48.80	47.19	75.66	0.99	120	289	5875	86,567	906	1116	17.25	21.25
LMP (Tang et al. 2017)	48.78	51.26	79.04	1.12	138	304	6654	86,245	481	595	9.13	11.29
TLMHT (Sheng et al. 2018a)	48.69	55.29	76.43	1.12	119	338	6632	86,504	413	642	7.86	12.22
STRN_MOT16 (Xu et al. 2019)	48.46	53.90	73.75	1.53	129	265	9038	84,178	747	2919	13.88	54.23
GCRA (Ma et al. 2018a)	48.16	48.55	77.50	0.86	98	312	5104	88,586	821	1117	15.97	21.73
FWT (Henschel et al. 2018)	47.77	44.28	75.51	1.50	145	290	8886	85,487	852	1534	16.04	28.88
MOTDT (Long et al. 2018)	47.63	50.94	74.81	1.56	115	291	9253	85,431	792	1858	14.90	34.96
NLLMPa (Levinkov et al. 2017)	47.58	47.34	78.51	0.99	129	307	5844	89,093	629	768	12.30	15.02
EAGS16 (Sheng et al. 2018b)	47.41	50.13	75.95	1.41	131	324	8369	86,931	575	913	10.99	17.45
JCSTD (Tian et al. 2019)	47.36	41.10	74.43	1.36	109	276	8076	86,638	1266	2697	24.12	51.39
ASTT (Tao et al. 2018)	47.24	44.27	76.08	0.79	124	316	4680	90,877	633	814	12.62	16.23
eHAF16 (Sheng et al. 2018c)	47.22	52.44	75.69	2.13	141	325	12,586	83,107	542	787	9.96	14.46
AMIR (Sadeghian et al. 2017)	47.17	46.29	75.82	0.45	106	316	2681	92,856	774	1675	15.77	34.14
JointMC (MCjoint) (Keuper et al. 2018)	47.10	52.26	76.27	1.13	155	356	6703	89,368	370	598	7.26	11.73
YOONKJI16 (Yoon et al. 2020)	46.96	50.05	75.76	1.33	125	317	7901	88,179	627	945	12.14	18.30
NOMT_16 (Choi 2015)	46.42	53.30	76.56	1.65	139	314	9753	87,565	359	504	6.91	9.70
JMC (Tang et al. 2016)	46.28	46.31	75.68	1.08	118	301	6373	90,914	657	1114	13.10	22.22
DD_TAMA16 (Yoon et al. 2019b)	46.20	49.43	75.42	0.87	107	334	5126	92,367	598	1127	12.12	22.84
DMAN_16 (Zhu et al. 2018)	46.08	54.82	73.77	1.34	132	324	7909	89,874	532	1616	10.49	31.87
STAM16 (Chu et al. 2017)	45.98	50.05	74.92	1.16	111	331	6895	91,117	473	1422	9.46	28.43
RAR16pub (Fang et al. 2018)	45.87	48.77	74.84	1.16	100	318	6871	91,173	648	1992	12.96	39.85
MHT_DAM_16 (Kim et al. 2015)	45.83	46.06	76.34	1.08	123	328	6412	91,758	590	781	11.88	15.72
MTDF (Fu et al. 2019)	45.72	40.07	72.63	2.03	107	276	12,018	84,970	1987	3377	37.21	63.24
INTERA_MOT (Lan et al. 2018)	45.40	47.66	74.41	2.27	137	294	13,407	85,547	600	930	11.30	17.52
EDMT (Chen et al. 2017a)	45.34	47.86	75.94	1.88	129	303	11,122	87,890	639	946	12.34	18.27
DCCRF16 (Zhou et al. 2018a)	44.76	39.67	75.63	0.95	107	321	5613	94,133	968	1378	20.01	28.49
TBSS (Zhou et al. 2018b)	44.58	42.64	75.18	0.70	93	333	4136	96,128	790	1419	16.71	30.01

Table 2 continued

Method	MOTA	IDF1	MOTP	FAR	MT	ML	FP	FN	IDSW	FM	IDSWR	FMR
OTCD_1_16 (Liu et al. 2019)	44.36	45.62	75.36	0.97	88	361	5759	94,927	759	1787	15.83	37.28
QuadMOT16 (Son et al. 2017)	44.10	38.27	76.40	1.08	111	341	6388	94775	745	1096	15.52	22.83
CDA_DDALv2 (Bae and Yoon 2018)	43.89	45.13	74.69	1.09	81	337	6450	95,175	676	1795	14.14	37.55
LFNF16 (Sheng et al. 2017)	43.61	41.62	76.63	1.12	101	347	6616	95,363	836	938	17.53	19.67
oICF_16 (Kieritz et al. 2016)	43.21	49.33	74.31	1.12	86	368	6651	96,515	381	1404	8.10	29.83
MHT_bLSTM6 (Kim et al. 2018)	42.10	47.84	75.85	1.97	113	337	11,637	93,172	753	1156	15.40	23.64
LINF1_16 (Fagot-Bouquet et al. 2016)	41.01	45.69	74.85	1.33	88	389	7896	99,224	430	963	9.43	21.13
PHD_GSDL16 (Fu et al. 2018)	41.00	43.14	75.90	1.10	86	315	6498	99,257	1810	3650	39.73	80.11
GMPHD_ReId Baisa (2019b)	40.42	49.71	75.25	1.11	85	329	6572	101,266	792	2529	17.81	56.88
AM_ADM (Lee et al. 2018)	40.12	43.79	75.45	1.44	54	351	8503	99,891	789	1736	17.45	38.40
EAMTT_pub (Sanchez-Matilla et al. 2016)	38.83	42.43	75.15	1.37	60	373	8114	102,452	965	1657	22.03	37.83
OVBT (Ban et al. 2016)	38.40	37.82	75.39	1.95	57	359	11,517	99,463	1321	2140	29.07	47.09
GMMCP (Dehghan et al. 2015)	38.10	35.50	75.84	1.12	65	386	6607	105,315	937	1669	22.18	39.51
LTTSC-CRF (Le et al. 2016)	37.59	42.06	75.94	2.02	73	419	11,969	101,343	481	1012	10.83	22.79
JCmin_MOT (Boragule and Jeon 2017)	36.65	36.16	75.86	0.50	57	413	2936	111,890	667	831	17.27	21.51
HISP_T (Baisa 2018)	35.87	28.93	76.07	1.08	59	380	6412	107,918	2594	2298	63.56	56.31
LP2D_16 (Leal-Taixé et al. 2014)	35.74	34.18	75.84	0.86	66	385	5084	111,163	915	1264	23.44	32.39
GM_PHD_DAL (Baisa 2019a)	35.13	26.58	76.59	0.40	53	390	2350	111,886	4047	5338	104.75	138.17
TBD_16 (Geiger et al. 2014)	33.74	0.00	76.53	0.98	55	411	5804	112,587	2418	2252	63.22	58.88
GM_PHD_N1T (Baisa and Wallace 2019)	33.25	25.47	76.84	0.30	42	425	1750	116,452	3499	3594	96.85	99.47
CEM_16 (Milan et al. 2014)	33.19	N/A	75.84	1.16	59	413	6837	114,322	642	731	17.21	19.60
GMPHD_HDA (Song and Jeon 2016)	30.52	33.37	75.42	0.87	35	453	5169	120,970	539	731	16.02	21.72
SMOT_16 (Dicle et al. 2013)	29.75	N/A	75.18	2.94	40	362	17,426	107,552	3108	4483	75.79	109.32
JPDA_m_16 (Rezatofighi et al. 2015)	26.17	N/A	76.34	0.62	31	512	3689	130,549	365	638	12.85	22.47
DP_NMS_16 (Pirsiavash et al. 2011)	26.17	31.19	76.34	0.62	31	512	3689	130,557	365	638	12.86	22.47

Performance of several trackers according to different metrics

Table 3 The *MOT17* leaderboard

Method	MOTA	IDF1	MOTP	FAR	MT	ML	FP	FN	IDSW	FM	IDSWR	FMR
MPNTrack (Brasó and Leal-Taixé 2020)	58.85	61.75	78.62	0.98	679	788	17,413	213,594	1185	2265	19.07	36.45
Tracktor++v2 (Bergmann et al. 2019)	56.35	55.12	78.82	0.50	498	831	8866	235,449	1987	3763	34.10	64.58
TrctrD17 (Xu et al. 2020)	53.72	53.77	77.23	0.66	458	861	11,731	247,447	1947	4792	34.68	85.35
Tracktor++ (Bergmann et al. 2019)	53.51	52.33	77.98	0.69	459	861	12,201	248,047	2072	4611	36.98	82.28
JBNOT (Henschel et al. 2019)	52.63	50.77	77.12	1.78	465	844	31,572	232,659	3050	3792	51.90	64.53
FAMNet (Chu and Ling 2019)	52.00	48.71	76.48	0.80	450	787	14,138	253,616	3072	5318	55.80	96.60
eTC17 (Wang et al. 2019)	51.93	58.13	76.34	2.04	544	836	36,164	232,783	2288	3071	38.95	52.28
eHAF17 (Sheng et al. 2018c)	51.82	54.72	77.03	1.87	551	893	33,212	236,772	1834	2739	31.60	47.19
YOONKJI17 (Yoon et al. 2020)	51.37	53.98	77.00	1.64	500	878	29,051	243,202	2118	3072	37.23	53.99
FWT_17 (Henschel et al. 2018)	51.32	47.56	77.00	1.36	505	830	24,101	247,921	2648	4279	47.24	76.33
NOTA (Chen et al. 2019)	51.27	54.46	76.68	1.13	403	833	20,148	252,531	2285	5798	41.36	104.95
JointMC (jCC) (Keuper et al. 2018)	51.16	54.50	75.92	1.46	493	872	25,937	247,822	1802	2984	32.13	53.21
STRN_MOT17 (Xu et al. 2019)	50.90	55.98	75.58	1.42	446	797	25,295	249,365	2397	9363	42.95	167.78
MOTDT17 (Long et al. 2018)	50.85	52.70	76.58	1.36	413	841	24,069	250,768	2474	5317	44.53	95.71
MHT_DAM_17 (Kim et al. 2015)	50.71	47.18	77.52	1.29	491	869	22,875	252,889	2314	2865	41.94	51.92
TLMHT_17 (Sheng et al. 2018a)	50.61	56.51	77.65	1.25	415	1022	22,213	255,030	1407	2079	25.68	37.94
EDMT17 (Chen et al. 2017a)	50.05	51.25	77.26	1.82	509	855	32,279	247,297	2264	3260	40.31	58.04
GMPHDOGM17 (Song et al. 2019)	49.94	47.15	77.01	1.35	464	895	24,024	255,277	3125	3540	57.07	64.65
MTDF17 (Fu et al. 2019)	49.58	45.22	75.48	2.09	444	779	37,124	241,768	5567	9260	97.41	162.03
PHD_GM (Sanchez-Matilla and Cavallaro 2019)	48.84	43.15	76.74	1.48	449	830	26,260	257,971	4407	6448	81.19	118.79

Table 3 continued

Method	MOTA	IDF1	MOTP	FAR	MT	ML	FP	FN	IDSW	FM	IDSWR	FMR
OTCD_1_17 (Liu et al. 2019)	48.57	47.90	76.91	1.04	382	970	18,499	268,204	3502	5588	66.75	106.51
HAM_SADF17 (Yoon et al. 2018a)	48.27	51.14	77.22	1.18	402	981	20,967	269,038	1871	3020	35.76	57.72
DMAN (Zhu et al. 2018)	48.24	55.69	75.69	1.48	454	902	26,218	263,608	2194	5378	41.18	100.94
AM_ADM17 (Lee et al. 2018)	48.11	52.07	76.69	1.41	316	934	25,061	265,495	2214	5027	41.82	94.95
PHD_GSDL17 (Fu et al. 2018)	48.04	49.63	77.15	1.31	402	838	23,199	265,954	3998	8886	75.63	168.09
MHT_bLSTM (Kim et al. 2018)	47.52	51.92	77.49	1.46	429	981	25,981	268,042	2069	3124	39.41	59.51
MASS (Karunasekera et al. 2019)	46.95	45.99	76.11	1.45	399	856	25,733	269,116	4478	11,994	85.62	229.31
GMPHD_Rd17 (Baisa 2019b)	46.83	54.06	76.41	2.17	464	784	38,452	257,678	3865	8097	71.14	149.03
IOU17 (Bochinski et al. 2017)	45.48	39.40	76.85	1.13	369	953	19,993	281,643	5988	7404	119.56	147.84
LM_NN_17 (Babaee et al. 2019)	45.13	43.17	78.93	0.61	348	1088	10,834	296,451	2286	2463	48.17	51.90
FPSN (Lee and Kim 2019)	44.91	48.43	76.61	1.90	388	844	33,757	269,952	7136	14,491	136.82	277.84
HISP_T17 (Baisa 2019c)	44.62	38.79	77.19	1.43	355	913	25,478	276,395	10,617	7487	208.12	146.76
GMPHD_DAL (Baisa 2019a)	44.40	36.23	77.42	1.08	350	927	19,170	283,380	11,137	13,900	223.74	279.25
SAS_MOT17 (Maksai and Fua 2019)	44.24	57.18	76.42	1.66	379	1044	29,473	283,611	1529	2644	30.74	53.16
GMPHD_SHA (Song and Jeon 2016)	43.72	39.17	76.53	1.46	276	1012	25,935	287,758	3838	5056	78.33	103.18
SORT17 (Bewley et al. 2016a)	43.14	39.84	77.77	1.60	295	997	28,398	287,582	4852	7127	98.96	145.36
EAMTT_17 (Sanchez-Matilla et al. 2016)	42.63	41.77	76.03	1.73	300	1006	30,711	288,474	4488	5720	91.83	117.04
GMPHD_N1Tr (Baisa and Wallace 2019)	42.12	33.87	77.66	1.03	280	1005	18,214	297,646	10,698	10,864	226.43	229.94
GMPHD_KCF (Kutschbach et al. 2017)	39.57	36.64	74.54	2.87	208	1019	50,903	284,228	5811	7414	117.10	149.40
GM_PHD (Eiselein et al. 2012)	36.36	33.92	76.20	1.34	97	1349	23,723	330,767	4607	11,317	111.34	273.51

Performance of several trackers according to different metrics
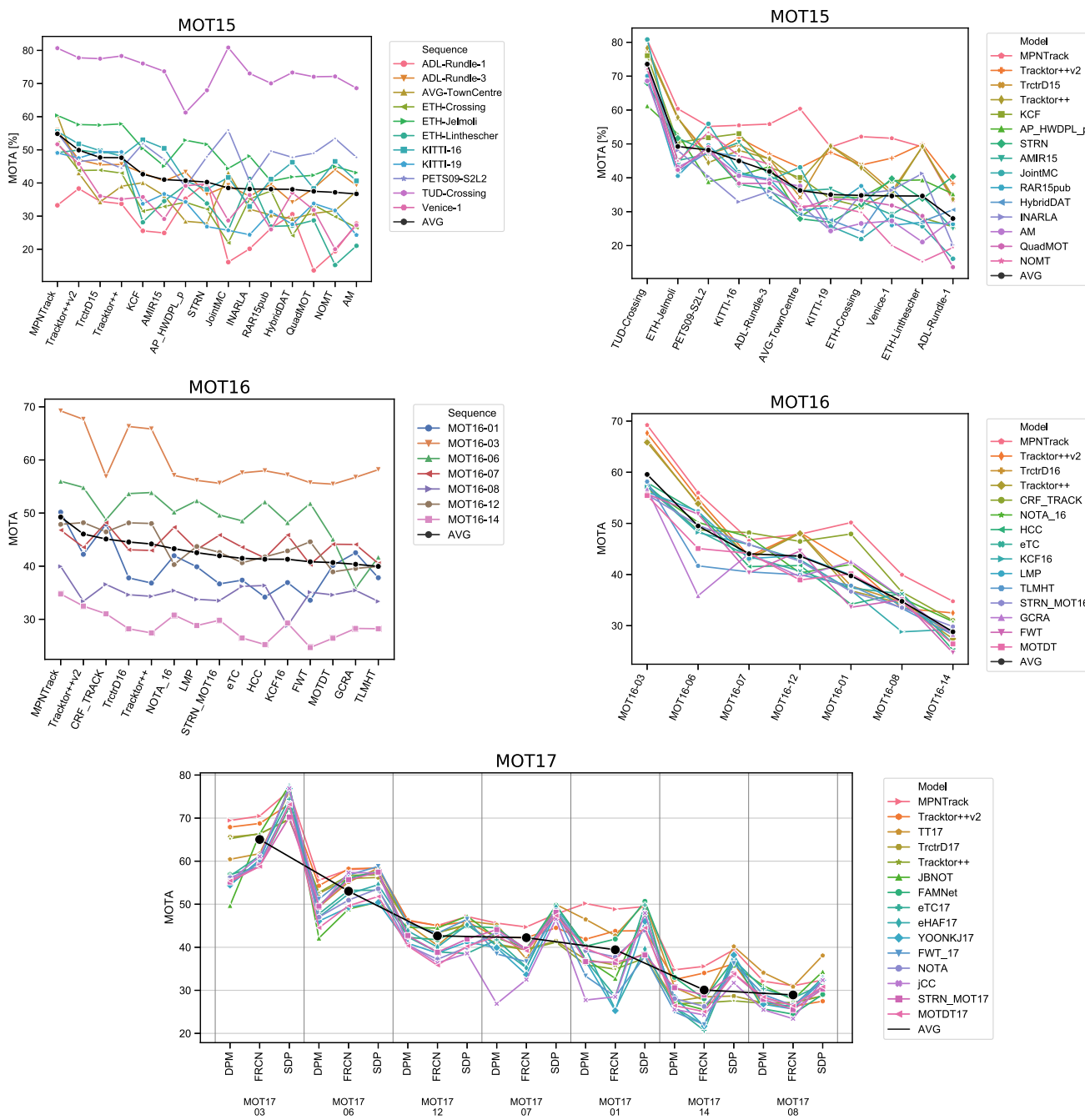
Fig. 5 Graphical overview of the top 15 trackers of all benchmarks. The entries are ordered from easiest sequence/best performing method, to hardest sequence/poorest performance, respectively. The mean performance across all sequences/submissions is depicted with a thick black line

MOT17. A small subset of the submissions[10] were done by the benchmark organizers and not by the original authors of the respective method. Results for *MOT15* are summarized in Table 1, for *MOT16* in Table 2 and for *MOT17* in Table 3. The performance of the top 15 ranked trackers is demonstrated in Fig. 5.

---

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

7.1 Trends in Tracking

Global optimization The community has long used the paradigm of tracking-by-detection for MOT, i.e., dividing the task into two steps: (i) object detection and (ii) data association, or temporal linking between detections. The data association problem could be viewed as finding a set of disjoint paths in a graph, where nodes in the graph represent object detections, and links hypothesize feasible associa-

tions. Detectors usually produce multiple spatially-adjacent detection hypotheses, that are usually pruned using heuristic non-maximum suppression (NMS).

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, e.g., a graphical model and solved with k-shortest paths in DP_NMS (Pirsiavash et al. 2011), as a linear program solved with the simplex algorithm in LP2D (Leal-Taixé et al. 2011), as a Conditional Random Field in DCO_X (Milan et al. 2016), SegTrack (Milan et al. 2015), LTTSC-CRF (Le et al. 2016), and GMMCP (Dehghan et al. 2015), using joint probabilistic data association filter (JPDA) (Rezatofighi et al. 2015) or as a variational Bayesian model in OVBT (Ban et al. 2016).

A number of tracking approaches investigate the efficacy of using a Probability Hypothesis Density (PHD) filter-based tracking framework (Baisa 2019a; Baisa 2019b; Baisa and Wallace 2019; Fu et al. 2018; Sanchez-Matilla et al. 2016; Song and Jeon 2016; Song et al. 2019; Wojke and Paulus 2016). This family of methods estimate states of multiple targets and data association simultaneously, reaching 30.72% MOTA on *MOT15* (GMPHD_OGM), 41% and 40.42% on *MOT16* (PHD_GSDL and GMPHD_ReId, respectively) and 49.94% (GMPHD_OGM) on *MOT17*.

Newer methods (Tang et al. 2015) bypassed the need to pre-process object detections with NMS. They proposed a multi-cut optimization framework, which finds the connected components in a graph that represent feasible solutions, clustering all detections that correspond to the same target. This family of methods (JMC (Tang et al. 2016), LMP (Tang et al. 2017), NLLMPA (Levinkov et al. 2017), JointMC (Keuper et al. 2018), HCC (Ma et al. 2018b)) achieve 35.65% MOTA on *MOT15* (JointMC), 48.78% and 49.25% (LMP and HCC, respectively) on *MOT16* and 51.16% (JointMC) on *MOT17*.

Motion Models A lot of attention has also been given to motion models, used as additional association affinity cues, e.g., SMOT (Dicle et al. 2013), CEM (Milan et al. 2014), TBD (Geiger et al. 2014), ELP (McLaughlin et al. 2015) and MotiCon (Leal-Taixé et al. 2014). The pairwise costs for matching two detections were based on either simple distances or simple appearance models, such as color histograms. These methods achieve around 38% MOTA on *MOT16* (see Table 2) and 25% on *MOT15* (see Table 1).

Hand-Crafted Affinity Measures After that, the attention shifted towards building robust pairwise similarity costs, mostly based on strong appearance cues or a combination of geometric and appearance cues. This shift is clearly reflected in an improvement in tracker performance and the ability for trackers to handle more complex scenarios. For example, LINF1 (Fagot-Bouquet et al. 2016) uses sparse appearance models, and oICF (Kieritz et al. 2016) use appearance models based on integral channel features. Top-performing methods

of this class incorporate long-term interest point trajectories, e.g., NOMT (Choi 2015), and, more recently, learned models for sparse feature matching JMC (Tang et al. 2016) and JointMC (Keuper et al. 2018) to improve pairwise affinity measures. As can be seen in Table 1, methods incorporating sparse flow or trajectories yielded a performance boost – in particular, NOMT is a top-performing method published in 2015, achieving MOTA of 33.67% on MOT15 and 46.42% on MOT16. Interestingly, the first methods outperforming NOMT on *MOT16* were published only in 2017 (AMIR (Sadeghian et al. 2017) and NLLMP (Levinkov et al. 2017)).

Towards Learning In 2015, we observed a clear trend towards utilizing learning to improve MOT. LP_SSVM (Wang and Fowlkes 2016) demonstrates a significant performance boost by learning the parameters of linear cost association functions within a network flow tracking framework, especially when compared to methods using a similar optimization framework but hand-crafted association cues, e.g. Leal-Taixé et al. (2014). The parameters are learned using structured SVM (Taskar et al. 2003). MDP (Xiang et al. 2015) goes one step further and proposes to learn track management policies (birth/death/association) by modeling object tracks as Markov Decision Processes (Thrun et al. 2005). Standard MOT evaluation measures (Stiefelhagen et al. 2006) are not differentiable. Therefore, this method relies on reinforcement learning to learn these policies. As can be seen in Table 1, this method outperforms the majority of methods published in 2015 by a large margin and surpasses 30% MOTA on *MOT15*.

In parallel, methods start leveraging the representational power of deep learning, initially by utilizing transfer learning. MHT_DAM (Kim et al. 2015) learns to adapt appearance models online using multi-output regularized least squares. Instead of weak appearance features, such as color histograms, they extract base features for each object detection using a pre-trained convolutional neural network. With the combination of the powerful MHT tracking framework (Reid 1979) and online-adapted features used for data association, this method surpasses MDP and attains over 32% MOTA on *MOT15* and 45% MOTA on *MOT16*. Alternatively, JMC (Tang et al. 2016) and JointMC (Keuper et al. 2018) use a pre-learned deep matching model to improve the pairwise affinity measures. All aforementioned methods leverage pre-trained models.

Learning Appearance Models The next clearly emerging trend goes in the direction of learning appearance models for data association in end-to-end fashion directly on the target (i.e., *MOT15*, *MOT16*, *MOT17*) datasets. SiameseCNN (Leal-Taixe et al. 2016) trains a siamese convolutional neural network to learn spatio-temporal embeddings based on object appearance and estimated optical flow using contrastive loss (Hadsell et al. 2006). The learned embeddings

Table 4 *MOT15, MOT16, MOT17* trackers and their characteristics

Method	Box–box affinity	App.	Opt.	Extra inputs	OA	TR	ON
MPNTrack (Brasó and Leal-Taixé 2020)	Appearance, geometry (L)	✓	MCF, LP	MC	✗	✓	✗
DeepMOT (Xu et al. 2020)	Re-id (L)	✓	–	MC	✗	✓	✓
TT17 (Zhang et al. 2020)	Appearance, geometry (L)	✓	MHT/MWIS	✗	✗	✗	✗
CRF_TRACK (Xiang et al. 2020)	Appearance, geometry (L)	✓	CRF	re-id	✗	✗	✓
Tracktor (Bergmann et al. 2019)	Re-id (L)	✓	–	MC, re-id	✓	✓	✓
KCF (Chu et al. 2019)	Re-id (L)	✓	Multicut	re-id	✗	✓	✓
STRN (Xu et al. 2019)	Geometry, appearance (L)	✓	Hungarian algorithm	–	✗	✗	✓
JBNOT (Henschel et al. 2019)	Joint, body distances	✗	Frank–Wolfe algorithm	Body joint det.	✗	✓	✗
FAMNet (Chu and Ling 2019)	(L)	✓	Rank-1 tensor approx.	–	✓	✗	✓
MHT_bLSTM (Kim et al. 2018)	Appearance, motion (L)	✓	MHT/MWIS	Pre-trained CNN	✗	✗	✓
JointMC (Keuper et al. 2018)	DeepMatching (L), geometric	✓	Multicut	OF, non-nms dets	✗	✗	✗
RAR (Fang et al. 2018)	Appearance, motion (L)	✓	Hungarian algorithm	–	✗	✗	✓
HCC (Ma et al. 2018b)	Re-id (L)	✓	Multicut	External re-id	○	✗	✗
FWT (Henschel et al. 2018)	DeepMatching, geometric	✓	Frank–Wolfe algorithm	Head detector	✗	✗	✗
DMAN (Zhu et al. 2018)	Appearance (L), geometry	✓	–	–	✓	✓	✓
eHAF (Zhu et al. 2018)	Appearance, motion	✓	MHT/MWIS	Super-pixels, OF	✗	✗	✗
QuadMOT (Son et al. 2017)	Re-id (L), motion	✓	Min-max label prop.	–	✓	✗	✓
STAM (Chu et al. 2017)	Appearance (L), motion	✓	–	–	✓	✓	✓
AMIR (Sadeghian et al. 2017)	Motion, appearance, interactions (L)	✓	Hungarian algorithm	–	✗	✗	✓
LMP (Tang et al. 2017)	Re-id (L)	✓	Multicut	Non-nms det., re-id	✓	✗	✗
NLLMPa (Levinkov et al. 2017)	DeepMatching	✓	Multicut	Non-NMS dets	✗	✗	✗
LP_SSVM (Wang and Fowlkes 2016)	Appearance, motion (L)	✓	MCF, greedy	–	✗	✗	✗
SiameseCNN (Leal-Taixe et al. 2016)	Appearance (L), geometry, motion	✓	MCF, LP	OF	✗	✗	✗
SCEA (Yoon et al. 2016)	Appearance, geometry	✓	Clustering	–	✗	✗	✓
JMC (Tang et al. 2016)	DeepMatching	✓	Multicut	Non-NMS dets	✗	✗	✗
LINF1 (Fagot-Bouquet et al. 2016)	Sparse representation	✓	MCMC	–	✗	✗	✗
EAMTTpub (Sanchez-Matilla et al. 2016)	2D distances	✗	Particle Filter	Non-NMS dets	✗	✗	✓
OVBT (Ban et al. 2016)	Dynamics from flow	✓	Variational EM	OF	✗	✗	✓
LTTSC-CRF (Le et al. 2016)	SURF	✓	CRF	SURF	✗	✗	✗
GMPHD_HDA (Song and Jeon 2016)	HoG similarity, color histogram	✓	GM-PHD filter	HoG	✗	✗	✓
DCO_X (Milan et al. 2016)	Motion, geometry	✓	CRF	–	✗	✗	✗
ELP (McLaughlin et al. 2015)	Motion	✗	MCF, LP	–	✗	✗	✗
GMMCP (Dehghan et al. 2015)	Appearance, motion	✓	GMMCP/CRF	–	✗	✗	✗
MDP (Xiang et al. 2015)	Motion (flow), geometry, appearance	✓	Hungarian algorithm	OF	✓	✗	✓
MHT_DAM (Kim et al. 2015)	(L)	✓	MHT/MWIS	–	✓	✗	✓

Table 4 continued

Method	Box–box affinity	App.	Opt.	Extra inputs	OA	TR	ON
NOMT (Choi 2015)	Interest point traj.	✓	CRF	OF	✗	✗	✗
JPDA_m (Rezatofighi et al. 2015)	Mahalanobis distance	✗	LP	–	✗	✗	✗
SegTrack (Milan et al. 2015)	Shape, geometry, motion	✓	CRF	OF, super-pixels	✗	✗	✗
TBD (Geiger et al. 2014)	IoU + NCC	✓	Hungarian algorithm	–	✗	✗	✗
CEM (Milan et al. 2014)	Motion	✗	Greedy sampling	–	✗	✗	✗
MotiCon (Leal-Taixé et al. 2014)	Motion descriptors	✗	MCF, LP	OF	✗	✗	✗
SMOT (Dicle et al. 2013)	Target dynamics	✗	Hankel Least Squares	–	✗	✗	✗
DP_NMS (Pirsiavash et al. 2011)	2D image distances	✗	k-shortest paths	–	✗	✗	✗
LP2D (Leal-Taixé et al. 2011)	2D image distances, IoU	✗	MCF, LP	–	✗	✗	✗

App. appearance model, *OA* online target appearance adaptation, *TR* target regression, *ON* online method, *(L)* learned. Components: *MC* motion compensation module, *OF* optical flow, *Re-id* learned re-identification module, *HoG* histogram of oriented gradients, *NCC* normalized cross-correlation, *IoU* intersection over union. **Association:** *GMMCP* Generalized maximum multi-clique problem, *MCF* Min-cost flow formulation (Zhang et al. 2008), *LP* linear programming, *MHT* multi-hypothesis tracking (Reid 1979), *MWIS* maximum independent set problem, *CRF* conditional random field formulation

are then combined with contextual cues for robust data association. This method uses similar linear programming based optimization framework (Zhang et al. 2008) compared to LP_SSVM (Wang and Fowlkes 2016), however, it surpasses it significantly performance-wise, reaching 29% MOTA on *MOT15*. This demonstrates the efficacy of fine-tuning appearance models directly on the target dataset and utilizing convolutional neural networks. This approach is taken a step further with QuadMOT (Son et al. 2017), which similarly learns spatio-temporal embeddings of object detections. However, they train their siamese network using quadruplet loss (Chen et al. 2017b) and learn to place embedding vectors of temporally-adjacent detections instances closer in the embedding space. These methods reach 33.42% MOTA in *MOT15* and 41.1% on *MOT16*.

The learning process, in this case, is supervised. Different from that, HCC (Ma et al. 2018b) learns appearance models in an unsupervised manner. To this end, they train their method using object trajectories obtained from the test set using offline correlation clustering-based tracking framework (Levinkov et al. 2017). TO (Manen et al. 2016), on the other hand, proposes to mine detection pairs over consecutive frames using single object trackers to learn affinity measures which are plugged into a network flow optimization tracking framework. Such methods have the potential to keep improving affinity models on datasets for which ground-truth labels are not available.

Online Appearance Model Adaptation The aforementioned methods only learn general appearance embedding vectors for object detection and do not adapt the tracking target appearance models online. Further performance is gained by methods that perform such adaptation online (Chu et al. 2017; Kim et al. 2015, 2018; Zhu et al. 2018). MHT_bLSTM (Kim et al. 2018) replaces the multi-output regularized least-squares learning framework of MHT_DAM (Kim et al. 2015) with a bi-linear LSTM and adapts both the appearance model as well the convolutional filters in an online fashion. STAM (Chu et al. 2017) and DMAN (Zhu et al. 2018) employ an ensemble of single-object trackers (SOTs) that share a convolutional backbone and learn to adapt the appearance model of the targets online during inference. They employ a spatio-temporal attention model that explicitly aims to prevent drifts in appearance models due to occlusions and interactions among the targets. Similarly, KCF (Chu et al. 2019) employs an ensemble of SOTs and updates the appearance model during tracking. To prevent drifts, they learn a tracking update policy using reinforcement learning. These methods achieve up to 38.9% MOTA on *MOT15*, 48.8% on MOT16 (KCF), and 50.71% on *MOT17* (MHT_DAM). Surprisingly, MHT_DAM out-performs its

bilinear-LSTM variant (MHT_bLSTM achieves a MOTA of 47.52%) on *MOT17*.

Learning to Combine Association Cues A number of methods go beyond learning only the appearance model. Instead, these approaches learn to encode and combine heterogeneous association cues. SiameseCNN (Leal-Taixe et al. 2016) uses gradient boosting to combine learned appearance embeddings with contextual features. AMIR (Sadeghian et al. 2017) leverages recurrent neural networks in order to encode appearance, motion, pedestrian interactions and learns to combine these sources of information. STRN (Xu et al. 2019) proposes to leverage relational neural networks to learn to combine association cues, such as appearance, motion, and geometry. RAR (Fang et al. 2018) proposes recurrent auto-regressive networks for learning a generative appearance and motion model for data association. These methods achieve 37.57% MOTA on *MOT15* and 47.17% on *MOT16*.

Fine-Grained Detection A number of methods employ additional fine-grained detectors and incorporate their outputs into affinity measures, e.g., a head detector in the case of FWT (Henschel et al. 2018), or a body joint detectors in JBNOT (Henschel et al. 2019), which are shown to help significantly with occlusions. The latter attains 52.63% MOTA on MOT17, which places it as the second-highest scoring method published in 2019.

Tracking-by-Regression Several methods leverage ensembles of (trainable) single-object trackers (SOTs), used to regress tracking targets from the detected objects, utilized in combination with simple track management (birth/death) strategies. We refer to this family of models as MOT-by-SOT or tracking-by-regression. We note that this paradigm for MOT departs from the traditional view of the multi-object tracking problem in computer vision as a generalized assignment problem (or multi-dimensional assignment problem), i.e. the problem of grouping object detections into a discrete set of tracks. Instead, methods based on target regression bring the focus back to the target state estimation. We believe the reasons for the success of these methods is two-fold: (i) rapid progress in learning-based SOT (Held et al. 2016; Li et al. 2018) that effectively leverages convolutional neural networks, and (ii) these methods can effectively utilize image evidence that is not covered by the given detection bounding boxes. Perhaps surprisingly, the most successful tracking-by-regression method, Tracktor (Bergmann et al. 2019), does not perform online appearance model updates (c.f., STAM, DMAN (Chu et al. 2017; Zhu et al. 2018) and KCF (Chu et al. 2019)). Instead, it simply re-purposes the regression head of the Faster R-CNN (Ren et al. 2015) detector, which is interpreted as the target regressor. This approach is most effective when combined with a motion compensation module and a learned re-identification module, attaining 46% MOTA on
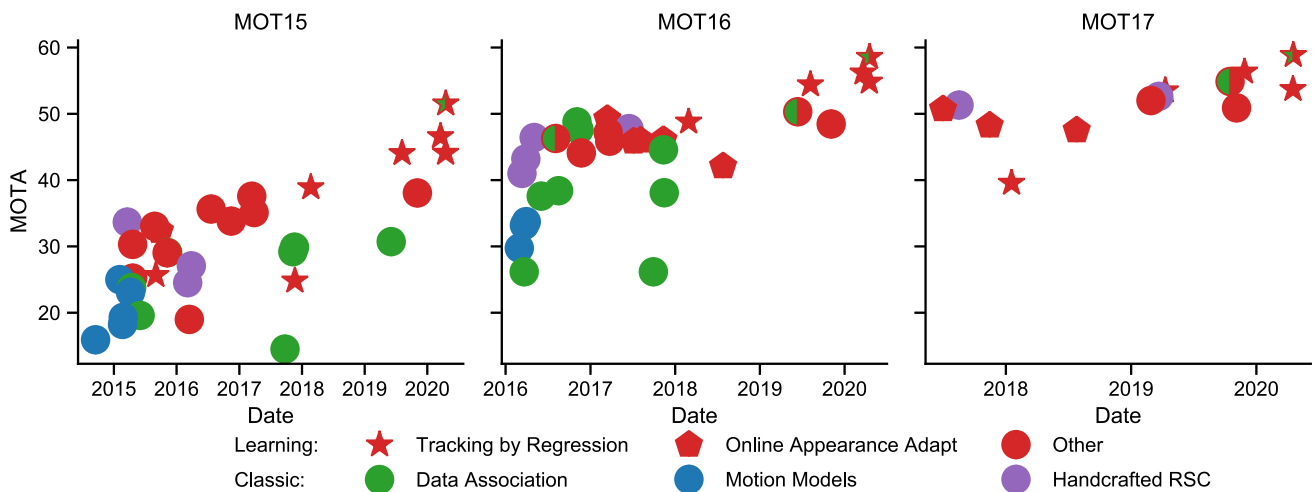
Fig. 6 Overview of tracker performances measured by their date of submission time and model type category
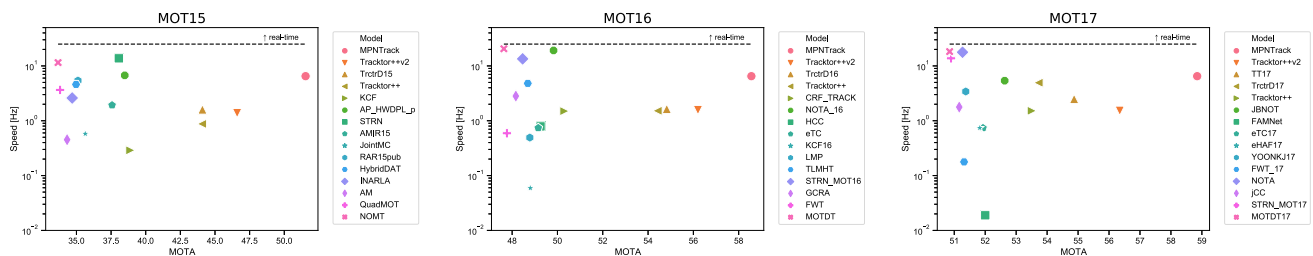
Fig. 7 Tracker performance measured by MOTA versus processing efficiency in frames per second for *MOT15*, *MOT16*, and *MOT17* on a log-scale. The latter is only indicative of the true value and has not been measured by the benchmark organizers. See text for details

MOT15 and 56% on MOT16 and MOT17, outperforming methods published in 2019 by a large margin.

Towards End-to-End Learning Even though tracking-by-regression methods brought substantial improvements, they are not able to cope with larger occlusions gaps. To combine the power of graph-based optimization methods with learning, MPNTrack (Brasó and Leal-Taixé 2020) proposes a method that leverages message-passing networks (Battaglia et al. 2016) to directly learn to perform data association via edge classification. By combining the regression capabilities of Tracktor (Bergmann et al. 2019) with a learned discrete neural solver, MPNTrack establishes a new state of the art, effectively using the best of both worlds—target regression and discrete data association. This method is the first one to surpass MOTA above 50% on *MOT15*. On the *MOT16* and *MOT17* it attains a MOTA of 58.56% and 58.85%, respectively. Nonetheless, this method is still not fully end-to-end trained, as it requires a projection step from the solution given by the graph neural network to the set of feasible solutions according to the network flow formulation and constraints.

Alternatively, (Xiang et al. 2020) uses MHT framework (Reid 1979) to link tracklets, while iteratively re-evaluating appearance/motion models based on progressively merged tracklets. This approach is one of the top on *MOT17*, achieving 54.87% MOTA.

In the spirit of combining optimization-based methods with learning, Zhang et al. (2020) revisits CRF-based tracking models and learns unary and pairwise potential functions in an end-to-end manner. On *MOT16*, this method attains MOTA of 50.31%.

We do observe trends towards learning to perform end-to-end MOT. To the best of our knowledge, the first method attempting this is RNN_LSTM (Milan et al. 2017), which jointly learns motion affinity costs and to perform bi-partite detection association using recurrent neural networks (RNNs). FAMNet (Chu and Ling 2019) uses a single network to extract appearance features from images, learns association affinities, and estimates multi-dimensional assignments of detections into object tracks. The multi-dimensional assignment is performed via a differentiable network layer that computes rank-1 estimation of the assignment tensor, which allows for back-propagation of the gradient. They perform learning with respect to binary cross-entropy loss between predicted assignments and ground-truth.

All aforementioned methods have one thing in common— they optimize network parameters with respect to proxy losses that do not directly reflect tracking quality, most com-

monly measured by the CLEAR-MOT evaluation measures (Stiefelhagen et al. 2006). To evaluate MOTA, the assignment between track predictions and ground truth needs to be established; this is usually performed using the Hungarian algorithm (Kuhn and Yaw 1955), which contains non-differentiable operations. To address this discrepancy DeepMOT (Xu et al. 2020) proposes the missing link—a differentiable matching layer that allows expressing a soft, differentiable variant of MOTA and MOTP.

Conclusion In summary, we observed that after an initial focus on developing algorithms for discrete data association (Dehghan et al. 2015; Le et al. 2016; Pirsiavash et al. 2011; Zhang et al. 2008), the focus shifted towards hand-crafting powerful affinity measures (Choi 2015; Kieritz et al. 2016; Leal-Taixé et al. 2014), followed by large improvements brought by learning powerful affinity models (Leal-Taixe et al. 2016; Son et al. 2017; Wang and Fowlkes 2016; Xiang et al. 2015).

In general, the major outstanding trends we observe in the past years all leverage the representational power of deep learning for learning association affinities, learning to adapt appearance models online (Chu et al. 2019, 2017; Kim et al. 2018; Zhu et al. 2018) and learning to regress tracking targets (Bergmann et al. 2019; Chu et al. 2019, 2017; Zhu et al. 2018). Figure 6 visualizes the promise of deep learning for tracking by plotting the performance of submitted models over time and by type.

The main common components of top-performing methods are: (i) learned single-target regressors (single-object trackers), such as (Held et al. 2016; Li et al. 2018), and (ii) re-identification modules (Bergmann et al. 2019). These methods fall short in bridging large occlusion gaps. To this end, we identified Graph Neural Network-based methods (Brasó and Leal-Taixé 2020) as a promising direction for future research. We observed the emergence of methods attempting to learn to track objects in end-to-end fashion instead of training individual modules of tracking pipelines (Chu and Ling 2019; Milan et al. 2017; Xu et al. 2020). We believe this is one of the key aspects to be addressed to further improve performance and expect to see more approaches leveraging deep learning for that purpose.

7.2 Runtime Analysis

Different methods require a varying amount of computational resources to track multiple targets. Some methods may require large amounts of memory while others need to be 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, regardless of the computational resources used. It is important to note that the resulting numbers are therefore only indicative of each approach and are not immediately comparable to one another.

Figure 7 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, the majority of methods are still far below real-time performance, which is assumed at 25 Hz. Second, the average processing rate $\sim$ 5 Hz does not differ much between the different sequences, which suggests that the different object densities (9 ped./fr. in *MOT15* and 26 ped./fr. in *MOT16/MOT17*) do not have a large impact on the speed the models. One explanation is that novel learning methods have an efficient forward computation, which does not vary much depending on the number of objects. This is in clear contrast to classic methods that relied on solving complex optimization problems at inference, which increased computation significantly as the pedestrian density increased. However, this conclusion has to be taken with caution because the runtimes are reported by the users on a trust base and cannot be verified by us.

7.3 Error Analysis

As we now, different applications have different requirements, e.g., for surveillance it is critical to have few false negatives, while for behavior analysis, having a false positive can mean computing wrong motion statistics. In this section, we take a closer look at the most common errors made by the tracking approaches. This simple analysis can guide researchers in choosing the best method for their task. In Fig. 8, 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 object detector, used as an input. A ratio below 1 indicates that the trackers have improved in terms of FN/FP over the detector. We show the performance of the top 15 trackers, averaged over sequences. We order them according to MOTA from left to right in decreasing order.

We observe all top-performing trackers reduce the amount of FPs and FNs compared to the public detections. While the trackers reduce FPs significantly, FNs are decreased only slightly. Moreover, we can see a direct correlation between the FN and tracker performance, especially for *MOT16* and *MOT17* datasets, since the number of FNs is much larger than the number of FPs. The question is then, why are methods not focusing on reducing FNs? It turns out that "filling the gaps" between detections, what is commonly thought trackers should do, is not an easy task.

It is not until 2018 that we see methods drastically decreasing the number of FNs, and as a consequence, MOTA performance leaps forward. As shown in Fig. 6, this is due to the appearance of learning-based tracking-by-regression methods (Bergmann et al. 2019; Brasó and Leal-Taixé 2020;
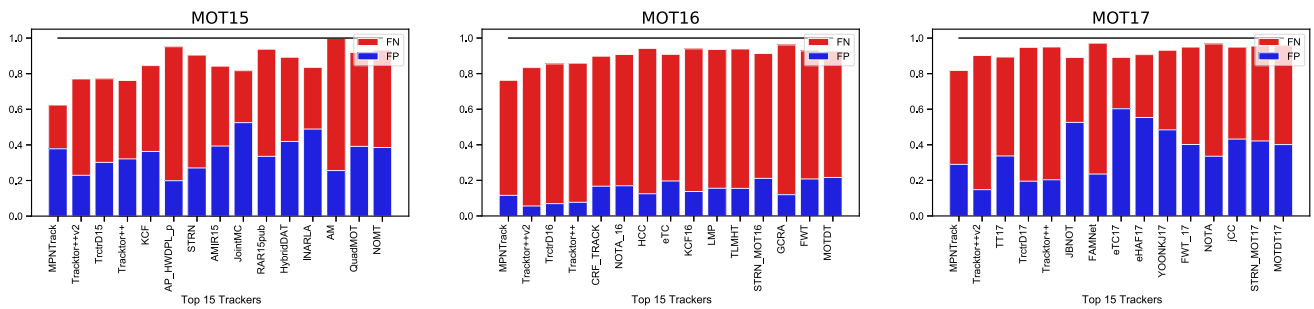
Fig. 8 Detailed error analysis. The plots show the error ratios for trackers w.r.t detector (taken at the lowest confidence threshold), for two types of errors: false positives (FP) and false negatives (FN). Values above 1 indicate a higher error count for trackers than for detectors. Note that most 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

Chu et al. 2017; Zhu et al. 2018). Such methods decrease the number of FNs the most by effectively using image evidence not covered by detection bounding boxes and regressing targets to areas where they are visible but missed by detectors. This brings us back to the common wisdom that trackers should be good at "filling the gaps" between detections.

Overall, it is clear that *MOT17* still presents a challenge both in terms of detection as well as tracking. It will require significant further future efforts to bring performance to the next level. In particular, the next challenge that future methods will need to tackle is bridging large occlusion gaps, which can not be naturally resolved by methods performing target regression, as these only work as long as the target is (partially) visible.

8 Conclusion and Future Work

We have introduced *MOTChallenge*, a standardized benchmark for a fair evaluation of single-camera multi-person tracking methods. We presented its first two data releases with about 35,000 frames of footage and almost 700,000 annotated pedestrians. Accurate annotations were carried out following a strict protocol, and extra classes such as vehicles, sitting people, reflections, or distractors were also annotated in the second release to provide further information to the community.

We have further analyzed the performance of 101 trackers; 73 *MOT15*, 74 *MOT16*, and 57 on *MOT17* obtaining several insights. In the past, at the center of vision-based MOT were methods focusing on global optimization for data association. Since then, we observed that large improvements were made by hand-crafting strong affinity measures and leveraging deep learning for learning appearance models, used for better data association. More recent methods moved towards directly regressing bounding boxes, and learning to adapt target appearance models online. As the most promising recent

trends that hold a large potential for future research, we identified the methods that are going in the direction of learning to track objects in an end-to-end fashion, combining optimization with learning.

We believe our Multiple Object Tracking Benchmark and the presented systematic analysis of existing tracking algorithms will help identify the strengths and weaknesses of the current state of the art and shed some light into promising future research directions.

Appendices

A Benchmark Submission

Our benchmark consists of the database and evaluation server on one hand, and the website as the user interface on the other. It is open to everyone who respects the submission policies (see next section). Before participating, every user is required to create an account, providing an institutional and not a generic e-mail address.[11]

After registering, the user can create a new tracker with a unique name and enter all additional details. It is mandatory to indicate:

- the full name and a brief description of the method
- a reference to the publication of the method, if already existing,
- whether the method operates online or on a batch of frames and whether the source code is publicly available,
- whether only the provided or also external training and detection data were used.

After creating all details of a new tracker, it is possible to assign open challenges to this tracker and submit results to the different benchmarks. To participate in a challenge the user has to provide the following information for each challenge they want to submit to:

- name of the challenge in which the tracker will be participating,
- a reference to the publication of the method, if already existing,
- the *total* runtime in seconds for computing the results for the test sequences and the hardware used, and
- whether only provided data was used for training, or also data from other sources were involved.

The user can then submit the results to the challenge in the format described in Sect. B.1. The tracking results are automatically evaluated and appear on the user's profile. The results are *not* automatically displayed in the public ranking table. The user can decide at any point in time to make the results public. Results can be published anonymously, e.g., to enable a blind review process for a corresponding paper. In this case, we ask to provide the venue and the paper ID or a similar unique reference. We request that a proper reference to the method's description is added upon acceptance of the paper. Anonymous entries are hidden from the benchmark after six months of inactivity.

---

[11] For accountability and to prevent abuse by using several email accounts.

The trackers and challenge meta information such as description, project page, runtime, or hardware can be edited at any time. Visual results of all public submissions, as well as annotations and detections, can be viewed and downloaded on the individual result pages of the corresponding tracker.

A.1 Submission Policy

The main goal of this benchmark is to provide a platform that allows for objective performance comparison of multiple target tracking approaches on real-world data. Therefore, we introduce a few simple guidelines that must be followed by all participants.

Training Ground truth is only provided for the training sequences. It is the participant's own responsibility to find the best setting using *only* the training data. The use of additional training data must be indicated during submission and will be visible in the public ranking table. The use of ground truth labels on the test data is strictly forbidden. This or any other misuse of the benchmark will lead to the deletion of the participant's account and their results.

Detections We also provide a unique set of detections (see Sect. 4.2) for each sequence. We expect all tracking-by-detection algorithms to use the given detections. In case the user wants to present results with another set of detections or is not using detections at all, this should be clearly stated during submission and will also be displayed in the results table.

Submission Frequency Generally, we expect one single submission for a particular method per benchmark. If for any reason the user needs to re-compute and re-submit the results (e.g. due to a bug discovered in the implementation), they may do so after a waiting period of 72 h after the last submission to submit to the same challenge with any of their trackers. This policy should discourage the use of the benchmark server for training and parameter tuning on the test data. The number of submissions is counted and displayed for each method. We allow a maximum number of 4 submissions per tracker and challenge. We allow a user to create several tracker instances for different tracking models. However, a user can only create a new tracker every 30 days. Under *no* circumstances must anyone create a second account and attempt to re-submit in order to bypass the waiting period. Such behavior will lead to the deletion of the accounts and exclusion of the user from participating in the benchmark.

A.2 Challenges and Workshops

We have two modalities for submission: the general open-end challenges and the special challenges. The main challenges, 2D MOT 2015, 3D MOT 2015, MOT16, and MOT17 are

always open for submission and are nowadays the standard evaluation platform for multi-target tracking methods submitting to computer vision conferences such as CVPR, ICCV or ECCV.

Special challenges are similar in spirit to the widely known PASCAL VOC series (Everingham et al. 2015), or the ImageNet competitions (Russakovsky et al. 2015). Each special challenge is linked to a workshop. The first edition of our series was the WACV 2015 Challenge that consisted of six outdoor sequences with both moving and static cameras, followed by the 2nd edition held in conjunction with ECCV 2016 on which we evaluated methods on the new *MOT16* sequences. The *MOT17* sequences were presented in the Joint Workshop on Tracking and Surveillance in conjunction with the Performance Evaluation of Tracking and Surveillance (PETS) (Ferryman and Ellis 2010; Ferryman and Shahrokni 2009) benchmark at the Conference on Vision and Pattern Recognition (CVPR) in 2017. The results and winning methods were presented during the respective workshops. Submission to those challenges is open only for a short period of time, i.e., there is a fixed submission deadline for all participants. Each method must have an accompanying paper presented at the workshop. The results of the methods are kept hidden until the date of the workshop itself when the winning method is revealed and a prize is awarded.

B MOT 15

We have compiled a total of 22 sequences, of which we use half for training and half for testing. The annotations of the testing sequences are not released in order to avoid (over)fitting of the methods to the specific sequences. Nonetheless, the test data contains over 10 minutes of footage and 61,440 annotated bounding boxes, therefore, it is hard for researchers to over-tune their algorithms on such a large amount of data. This is one of the major strengths of the benchmark. We classify the sequences according to:

We classify the sequences according to:

- *Moving or static camera* the camera can be held by a person, placed on a stroller (Ess et al. 2008) or on a car (Geiger et al. 2012), or can be positioned fixed in the scene.
- *Viewpoint* the camera can overlook the scene from a high position, a medium position (at pedestrian's height), or at a low position.
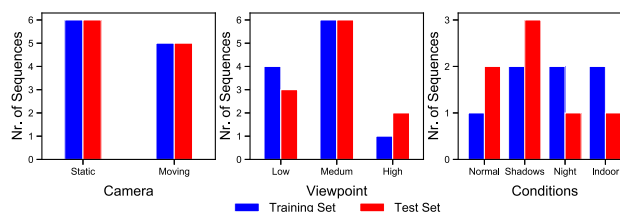
Fig. 9 Comparison histogram between training and testing sequences of *static* versus *moving* camera, camera *viewpoint*: low, medium or high, *conditions*: normal, shadows, night or indoor

- *Weather* the illumination conditions in which the sequence was taken. Sequences with strong shadows and saturated parts of the image make tracking challenging, while night sequences contain a lot of motion blur, which is often a problem for detectors. Indoor sequences contain a lot of reflections, while the sequences classified as normal do not contain heavy illumination artifacts that potentially affect tracking.

We divide the sequences into training and testing to have a balanced distribution, as shown in Fig. 9.

B.1 Data Format

All images were converted to JPEG and named sequentially to a 6-digit file name (e.g. 000001.jpg). Detection and annotation files are simple comma-separated value (CSV) files. Each line represents one object instance, and it contains 10 values as shown in Table 6.

The first number indicates in which frame the object appears, while the second number identifies that object as belonging to a trajectory by assigning a unique ID (set to −1 in a detection file, as no ID is assigned yet). Each object can be assigned to only one trajectory. The next four numbers indicate the position of the bounding box of the pedestrian in 2D image coordinates. The position is indicated by the top-left corner as well as the width and height of the bounding box. This is followed by a single number, which in the case of detections denotes their confidence score. The last three numbers indicate the 3D position in real-world coordinates of the pedestrian. This position represents the feet of the person. In the case of 2D tracking, these values will be ignored and can be left at −1.

Table 5 Overview of the sequences currently included in the *MOT15* benchmark

Name	FPS	Resolution	Length	Tracks	Boxes	Density	3D	Camera	Viewpoint	Conditions
Training sequences										
TUD-Stadtmitte (Andriluka et al. 2010)	25	640 × 480	179 (00:07)	10	1156	6.5	Yes	Static	Medium	Normal
TUD-Campus (Andriluka et al. 2010)	25	640 × 480	71 (00:03)	8	359	5.1	No	Static	Medium	Normal
PETS09-S2L1 (Ferryman and Ellis 2010)	7	768 × 576	795 (01:54)	19	4476	5.6	Yes	Static	High	Normal
ETH-Bahnhof (Ess et al. 2008)	14	640 × 480	1000 (01:11)	171	5415	5.4	Yes	Moving	Low	Normal
ETH-Sunnyday(Ess et al. 2008)	14	640 × 480	354 (00:25)	30	1858	5.2	Yes	Moving	Low	Shadows
ETH-Pedcross2(Ess et al. 2008)	14	640 × 480	840 (01:00)	133	6263	7.5	No	Moving	Low	Shadows
ADL-Rundle-6 (new)	30	1920 × 1080	525 (00:18)	24	5009	9.5	No	Static	Low	Indoor
ADL-Rundle-8 (new)	30	1920 × 1080	654 (00:22)	28	6783	10.4	No	Moving	Medium	Night
KITTI-13 (Geiger et al. 2012)	10	1242 × 375	340 (00:34)	42	762	2.2	No	Moving	Medium	Shadows
KITTI-17 (Geiger et al. 2012)	10	1242 × 370	145 (00:15)	9	683	4.7	No	Static	Medium	Shadows
Venice-2 (new)	30	1920 × 1080	600 (00:20)	26	7141	11.9	No	Static	Medium	Normal
Total training			5503 (06:29)	500	39,905	7.3				
Testing sequences										
TUD-Crossing (Andriluka et al. 2018)	25	640 × 480	201 (00:08)	13	1102	5.5	No	Static	Medium	Normal
PETS09-S2L2 (Ferryman and Ellis 2010)	7	768 × 576	436 (01:02)	42	9641	22.1	Yes	Static	High	Normal
ETH-Jelmoli (Ess et al. 2008)	14	640 × 480	440 (00:31)	45	2537	5.8	Yes	Moving	Low	Shadows
ETH-Linthescher (Ess et al. 2008)	14	640 × 480	1194 (01:25)	197	8930	7.5	Yes	Moving	Low	Shadows
ETH-Crossing (Ess et al. 2008)	14	640 × 480	219 (00:16)	26	1003	4.6	No	Moving	Low	Normal
AVG-TownCentre (Benfold and Reid 2011)	2.5	1920 × 1080	450 (03:45)	226	7148	15.9	Yes	Static	High	Normal
ADL-Rundle-1 (new)	30	1920 × 1080	500 (00:17)	32	9306	18.6	No	Moving	Medium	Normal
ADL-Rundle-3 (new)	30	1920 × 1080	625 (00:21)	44	10166	16.3	No	Static	Medium	Shadows
KITTI-16 (Geiger et al. 2012)	10	1242 × 370	209 (00:21)	17	1701	8.1	No	Static	Medium	Shadows
KITTI-19 (Geiger et al. 2012)	10	1242 × 374	1059 (01:46)	62	5343	5.0	No	Moving	Medium	Shadows
Venice-1 (new)	30	1920 × 1080	450 (00:15)	17	4563	10.1	No	Static	Medium	Normal
Total testing			5783 (10:07)	721	61,440	10.6				

An example of such a detection 2D file is:

```
1, -1, 794.2, 47.5, 71.2, 174.8, 67.5, -1, -1,
                -1
1, -1, 164.1, 19.6, 66.5, 163.2, 29.4, -1, -1,
                -1
1, -1, 875.4, 39.9, 25.3, 145.0, 19.6, -1, -1,
                -1
2, -1, 781.7, 25.1, 69.2, 170.2, 58.1, -1, -1,
                -1
```

For the ground truth and results files, the 7th value (confidence score) acts as a flag whether the entry is to be considered. A value of 0 means that this particular instance is ignored in the evaluation, while a value of 1 is used to mark it as active. An example of such an annotation 2D file is:

```
1, 1, 794.2, 47.5, 71.2, 174.8, 1, -1, -1, -1
1, 2, 164.1, 19.6, 66.5, 163.2, 1, -1, -1, -1
1, 3, 875.4, 39.9, 25.3,  35.0, 0, -1, -1, -1
2, 1, 781.7, 25.1, 69.2, 170.2, 1, -1, -1, -1
```

In this case, there are 2 pedestrians in the first frame of the sequence, with identity tags 1, 2. The third pedestrian is too small and therefore not considered, which is indicated with a flag value (7th value) of 0. In the second frame, we can see that pedestrian 1 remains in the scene. Note, that since this is a 2D annotation file, the 3D positions of the pedestrians are ignored and therefore are set to -1. All values including the bounding box are 1-based, i.e. the top left corner corresponds to (1, 1).

To obtain a valid result for the entire benchmark, a separate CSV file following the format described above must be created for each sequence and called "Sequence-Name.txt". All files must be compressed into a single zip file that can then be uploaded to be evaluated.

C MOT16 and MOT17 Release

Table 9 presents an overview of the *MOT16* and *MOT17* dataset.

C.1 Annotation Rules

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 this section, we define a clear protocol that was obeyed throughout the entire dataset annotations of *MOT16* and *MOT17* to guarantee consistency.

C.1.1 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
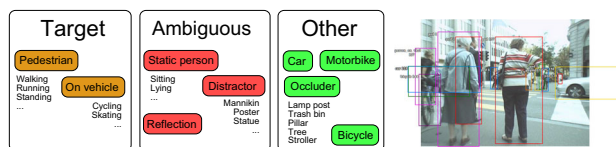
Fig. 10 Left: An overview of annotated classes. The classes in orange will be the central ones to evaluate on. The red classes include ambiguous cases such that neither recovering nor missing will be penalized in the evaluation. The classes in green are annotated for training purposes and for computing 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)

will be the center of attention of our annotations. We divide the pertinent classes into three categories:

(i) *moving* or *standing* pedestrians;
(ii) people that are *not in an upright position* or artificial representations of humans; and
(iii) *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 fire a detection response, such as mannikins, 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. These annotations will not play any role in the evaluation, but 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 Table 7, and in Fig. 10 we present a diagram of the classes of objects we annotate, as well as a sample frame with annotations.

Table 6 Data format for the input and output files, both for detection and annotation files

Position	Name	Description
1	Frame number	Indicate at which frame the object is present
2	Identity number	Each pedestrian trajectory is identified by a unique ID (-1 for detections)
3	Bounding box left	Coordinate of the top-left corner of the pedestrian bounding box
4	Bounding box top	Coordinate of the top-left corner of the pedestrian bounding box
5	Bounding box width	Width in pixels of the pedestrian bounding box
6	Bounding box height	Height in pixels of the pedestrian bounding box
7	Confidence score	Indicates how confident the detector is that this instance is a pedestrian. For the ground truth and results, it acts as a flag whether the entry is to be considered.
8	x	3D x position of the pedestrian in real-world coordinates (-1 if not available)
9	y	3D y position of the pedestrian in real-world coordinates (-1 if not available)
10	z	3D z position of the pedestrian in real-world coordinates (-1 if not available)

Table 7 Annotation rules

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 other 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

C.1.2 Bounding Box Alignment

The bounding box is aligned with the object's extent as accurately as possible. It should contain all object pixels belonging to that instance and at the same time be as tight as possible. This implies 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 object), then several boxes may be used to better approximate the extent of that object.

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

C.1.3 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 ≈ 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. A person leaving the field of view and re-appearing at a later point is assigned a new ID.

C.1.4 Minimal Size

Although the evaluation will only take into account pedestrians that have a minimum height in pixels, annotations contain all objects of all sizes as long as they are distinguishable by the annotator. In other words, all targets are annotated independently of their sizes in the image.

C.1.5 Occlusions

There is no need to explicitly annotate the level of occlusion. This value is be computed automatically using the annotations. We leverage the assumption that for two or more overlapping bounding boxes the object with the lowest y-value of the bounding box is closest to the camera and therefore occlude the other object behind it. Each target is fully annotated through occlusions as long as its extent and location can be determined accurately. If a target becomes completely occluded in the middle of a sequence and does not become visible later, the track is terminated (marked as 'outside of view'). If a target reappears after a prolonged period such that its location is ambiguous during the occlusion, it is assigned a new ID.

C.1.6 Sanity Check

Upon annotating all sequences, a "sanity check" is carried out to ensure that no relevant entities are missed. To that end, we run a pedestrian detector on all videos and add all high-confidence detections that correspond to either humans or distractors to the annotation list.

C.2 Data Format

All images were converted to JPEG and named sequentially to a 6-digit file name (e.g. 000001.jpg). Detection and annotation files are simple comma-separated value (CSV) files. Each line represents one object instance and contains 9 values as shown in Table 11.

The first number indicates in which frame the object appears, while the second number identifies that object as
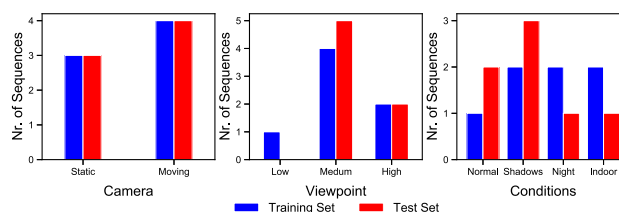
Fig. 11 Comparison histogram between training and testing sequences of *MOT16/MOT17*: camera: static vs. moving camera, viewpoint: low, medium or high, conditions: normal, shadows, night or indoor

belonging to a trajectory by assigning a unique ID (set to −1 in a detection file, as no ID is assigned yet). Each object can be assigned to only one trajectory. The next four numbers indicate the position of the bounding box of the pedestrian in 2D image coordinates. The position is indicated by the top-left corner as well as the width and height of the bounding box. This is followed by a single number, which in the case of detections denotes their confidence score. The last two numbers for detection files are ignored (set to -1). An example of such a 2D detection file is:

```
1, -1, 794.2, 47.5, 71.2, 174.8, 67.5, -1, -1
1, -1, 164.1, 19.6, 66.5, 163.2, 29.4, -1, -1
1, -1, 875.4, 39.9, 25.3, 145.0, 19.6, -1, -1
2, -1, 781.7, 25.1, 69.2, 170.2, 58.1, -1, -1
```

For the ground truth and result files, the 7th value (confidence score) acts as a flag whether the entry is to be considered. A value of 0 means that this particular instance is ignored in the evaluation, while a value of 1 is used to mark it as active. The 8th number indicates the type of object annotated, following the convention of Table 12. The last number shows the visibility ratio of each bounding box. This can be due to occlusion by another static or moving object, or to image border cropping. An example of such an annotation 2D file is:

```
1, 1, 794.2, 47.5, 71.2, 174.8, 1, 1, 0.8
1, 2, 164.1, 19.6, 66.5, 163.2, 1, 1, 0.5
2, 4, 781.7, 25.1, 69.2, 170.2, 0, 12, 1.
```

In this case, there are 2 pedestrians in the first frame of the sequence, with identity tags 1, 2. In the second frame, we can see a reflection (class 12), which is to be considered by the evaluation script and will neither count as a false negative nor as a true positive, independent of whether it is correctly recovered or not. All values including the bounding box are 1-based, i.e. the top left corner corresponds to (1, 1).

To obtain a valid result for the entire benchmark, a separate CSV file following the format described above must be created for each sequence and called "Sequence-Name.txt". All files must be compressed into a single ZIP file that can then be uploaded to be evaluated.

Table 8 Overview of the types of annotations currently found in the *MOT16/MOT17* benchmark

Sequence	Pedestrian	Person on vehicle	Car	Bicycle	Motorbike	Vehicle (non-mot.)	Static person	Distractor	Occluder (ground)	Occluder (full)	Refl.	Total
MOT16/17-01	6395/6450	346	0/0	341	0	0	4790/5230	900	3150/4050	0	0/0	15,922/17,317
MOT16/17-02	17,833/18,581	1549	0/0	1559	0	0	5271/5271	1200	1781/1843	0	0/0	29,193/30,003
MOT16/17-03	104,556/104,675	70	1500/1500	12,060	1500	0	6000/6000	0	24,000/24,000	13,500	0/0	163,186/163,305
MOT16/17-04	47,557/47,557	0	1050/1050	11,550	1050	0	4798/4798	0	23,100/23,100	18,900	0/0	108,005/108,005
MOT16/17-05	6818/6917	315	196/196	315	0	11	0/0	16	0/235	0	0/0	7671/8013
MOT16/17-06	11,538/11,784	150	0/0	118	0	0	269/269	238	109/109	0	0/299	12,422/12,729
MOT16/17-07	16,322/16,893	0	0/0	0	0	0	2,023/2,023	0	1920/2420	0	0/131	20,265/21,504
MOT16/17-08	16,737/21,124	0	0/0	0	0	0	1715/3535	2719	6875/6875	0	0/0	28,046/34,253
MOT16/17-09	5257/5325	0	0/0	0	0	0	0/514	1575	1050/1050	0	948/1947	8830/10,411
MOT16/17-10	12,318/12,839	0	25/25	0	0	0	1376/1376	470	2740/2740	0	0/0	16,929/17,450
MOT16/17-11	9174/9436	0	0/0	0	0	0	0/82	306	596/596	0	0/181	10,076/10,617
MOT16/17-12	8295/8667	0	0/0	0	0	0	1012/1036	763	1394/1710	0	0/953	11,464/13,272
MOT16/17-13	11,450/11,642	0	4484/4918	103	0	0	0/0	4	2542/2733	680	0/122	19,263/20,202
MOT16/17-14	18,483/18,483	0	1563/1563	0	0	0	712/712	47	4062/4062	393	0/0	25,260/25,294
Total	292,733/300,373	2430	8818/9252	26,046	2550	11	27,966/30,846	8238	73,319/75,523	33,473	948/3633	476,532/492,375

Table 9 Overview of the sequences currently included in the *MOT16/MOT17* benchmark

Name	FPS	Resolution	Length	Tracks	Boxes	Density	Camera	Viewpoint	Conditions
Training sequences									
MOT16/17-02 (new)	30	1920 × 1080	600 (00:20)	54/62	17,833/18,581	29.7/31.0	Static	Medium	Cloudy
MOT16/17-04 (new)	30	1920 × 1080	1050 (00:35)	83/83	47,557/47557	45.3/45.3	Static	High	Night
MOT16/17-05 (Ess et al. 2008)	14	640 × 480	837 (01:00)	125/133	6818/6917	8.1/8.3	Moving	Medium	Sunny
MOT16/17-09 (new)	30	1920 × 1080	525 (00:18)	25/26	5257/5325	10.0/10.1	Static	Low	Indoor
MOT16/17-10 (new)	30	1920 × 1080	654 (00:22)	54/57	12,318/12,839	18.8/19.6	Moving	Medium	Night
MOT16/17-11 (new)	30	1920 × 1080	900 (00:30)	69/75	9174/9436	10.2/10.5	Moving	Medium	Mndoor
MOT16/17-13 (new)	25	1920 × 1080	750 (00:30)	107/110	11,450/11,642	15.3/15.5	Moving	High	Sunny
Total training			5316 (03:35)	517/546	110,407/112,297	20.8/21.1			
Testing sequences									
MOT16/17-01 (new)	30	1920 × 1080	450 (00:15)	23/24	6,395/6,450	14.2/14.3	Static	Medium	Cloudy
MOT16/17-03 (new)	30	1920 × 1080	1,500 (00:50)	148/148	104,556/104,675	69.7/69.8	Static	High	Night
MOT16/17-06 (Ess et al. 2008)	14	640 × 480	1,194 (01:25)	221/222	11,538/11,784	9.7/9.9	Moving	Medium	Sunny
MOT16/17-07 (new)	30	1920 × 1080	500 (00:17)	54/60	16,322/16,893	32.6/33.8	Moving	Medium	Shadow
MOT16/17-08 (new)	30	1920 × 1080	625 (00:21)	63/76	16,737/21,124	26.8/33.8	Static	Medium	Sunny
MOT16/17-12 (new)	30	1920 × 1080	900 (00:30)	86/91	8,295/8,667	9.2/9.6	Moving	Medium	Indoor
MOT16/17-14 (new)	25	1920 × 1080	750 (00:30)	164/164	18,483/18,483	24.6/24.6	Moving	High	Sunny
Total testing			5919 (04:08)	759/785	182,326/188,076	30.8/31.8			

Table 10 Detection bounding box statistics

| Seq | MOT16 | | MOT17 | | | | | |
| | DPM | | DPM | | FRCNN | | SDP | |
	nDet.	nDet./fr.	nDet.	nDet./fr.	nDet.	nDet./fr.	nDet.	nDet./fr.
MOT16/17-01	3775	8.39	3775	8.39	5514	12.25	5837	12.97
MOT16/17-02	7267	12.11	7267	12.11	8186	13.64	11,639	19.40
MOT16/17-03	85,854	57.24	85,854	57.24	65,739	43.83	80,241	53.49
MOT16/17-04	39,437	37.56	39,437	37.56	28,406	27.05	37,150	35.38
MOT16/17-05	4333	5.20	4333	5.20	3848	4.60	4767	5.70
MOT16/17-06	7851	6.58	7851	6.58	7809	6.54	8283	6.94
MOT16/17-07	11,309	22.62	11,309	22.62	9377	18.75	10,273	20.55
MOT16/17-08	10,042	16.07	10,042	16.07	6921	11.07	8118	12.99
MOT16/17-09	5976	11.38	5976	11.38	3049	5.81	3607	6.87
MOT16/17-10	8832	13.50	8832	13.50	9701	14.83	10,371	15.86
MOT16/17-11	8590	9.54	8590	9.54	6007	6.67	7509	8.34
MOT16/17-12	7764	8.74	7764	8.74	4726	5.32	5440	6.09
MOT16/17-13	5355	7.22	5355	7.22	8442	11.26	7744	10.41
MOT16/17-14	8781	11.71	8781	11.71	10,055	13.41	10,461	13.95
Total	215,166	19.19	215,166	19.19	177,780	15.84	211,440	18.84

Table 11 Data format for the input and output files, both for detection (DET) and annotation/ground truth (GT) files

Position	Name	Description
1	Frame number	Indicate at which frame the object is present
2	Identity number	Each pedestrian trajectory is identified by a unique ID (-1 for detections)
3	Bounding box left	Coordinate of the top-left corner of the pedestrian bounding box
4	Bounding box top	Coordinate of the top-left corner of the pedestrian bounding box
5	Bounding box width	Width in pixels of the pedestrian bounding box
6	Bounding box height	Height in pixels of the pedestrian bounding box
7	Confidence score	DET: Indicates how confident the detector is that this instance is a pedestrian. GT: It acts as a flag whether the entry is to be considered (1) or ignored (0).
8	Class	GT: Indicates the type of object annotated
9	Visibility	GT: Visibility ratio, a number between 0 and 1 that says how much of that object is visible. Can be due to occlusion and due to image border cropping

D Implementation Details of the Evaluation

In this section, we detail how to compute false positives, false negatives, and identity switches, which are the basic units for the evaluation metrics presented in the main paper. We also explain how the evaluation deals with special non-target cases: people behind a window or sitting people.

D.1 Tracker-to-Target Assignment

There are two common prerequisites for quantifying the performance of a tracker. One is to determine for each hypothesized output, whether it is a true positive (TP) that describes an actual (annotated) target, or whether the output is a false alarm (or false positive, FP). This decision is typically made by thresholding based on a defined distance (or dissimilarity) measure d between the coordinates of the true and predicted box placed around a target (see Sect. D.2). A target that is missed by any hypothesis is a false negative

Table 12 Label classes present in the annotation files and ID appearing in the 7th column of the files as described in Table 11

Label	ID
Pedestrian	1
Person on vehicle	2
Car	3
Bicycle	4
Motorbike	5
Non motorized vehicle	6
Static person	7
Distractor	8
Occluder	9
Occluder on the ground	10
Occluder full	11
Reflection	12

(FN). A good result is expected to have as few FPs and FNs as possible. Next to the absolute numbers, we also show the false positive ratio measured by the number of false alarms per frame (FAF), sometimes also referred to as false positives per image (FPPI) in the object detection literature.

The same target may be covered by multiple outputs. The second prerequisite before computing the numbers is then to establish the correspondence between all annotated and hypothesized objects under the constraint that a true object should be recovered at most once, and that one hypothesis cannot account for more than one target.

For the following, we assume that each ground-truth trajectory has one unique start and one unique endpoint, i.e., that it is not fragmented. Note that the current evaluation procedure does not explicitly handle target re-identification. In other words, when a target leaves the field-of-view and then reappears, it is treated as an unseen target with a new ID. As proposed in Stiefelhagen et al. (2006), the optimal matching is found using Munkres (a.k.a. Hungarian) algorithm. However, dealing with video data, this matching is not performed independently for each frame, but rather considering a temporal correspondence. More precisely, if a ground-truth object i is matched to hypothesis j at time $t-1$ *and* the distance (or dissimilarity) between i and j in frame t is below t_d, then the correspondence between i and j is carried over to frame t even if there exists another hypothesis that is closer to the actual target. A mismatch error (or equivalently an identity switch, IDSW) is counted if a ground-truth target i is matched to track j and the last known assignment was $k \neq j$. Note that this definition of ID switches is more similar to (Li et al. 2009) and stricter than the original one (Stiefelhagen et al. 2006). Also note that, while it is certainly desirable to keep the number of ID switches low, their absolute number alone is not always expressive to assess the overall performance, but should rather be considered concerning the number of recovered targets. The intuition is that a method that finds twice as

many trajectories will almost certainly produce more identity switches. For that reason, we also state the relative number of ID switches, which is computed as IDSW / Recall.

These relationships are illustrated in Fig. 12. For simplicity, we plot ground-truth trajectories with dashed curves, and the tracker output with solid ones, where the color represents a unique target ID. The grey areas indicate the matching threshold (see Sect. D.3). Each true target that has been successfully recovered in one particular frame is represented with a filled black dot with a stroke color corresponding to its matched hypothesis. False positives and false negatives are plotted as empty circles. See figure caption for more details.

After determining true matches and establishing correspondences it is now possible to compute the metrics. We do so by concatenating all test sequences and evaluating the entire benchmark. This is in general more meaningful than averaging per-sequences figures because of the large variation on the number of targets per sequence.

D.2 Distance Measure

The relationship between ground-truth objects and a tracker output is established using bounding boxes on the image plane. Similar to object detection (Everingham et al. 2015), the intersection over union (a.k.a. the Jaccard index) is usually employed as the similarity criterion, while the threshold t_d is set to 0.5 or 50%.

D.3 Target-Like Annotations

People are a common object class present in many scenes, but should we track all people in our benchmark? For example, should we track static people sitting on a bench? Or people on bicycles? How about people behind a glass? We define the target class of *MOT16* and *MOT17* as *all upright people, standing or walking, that are reachable along the viewing ray without a physical obstacle.* For instance, reflections or people behind a transparent wall or window are excluded. We also exclude from our target class people on bicycles (riders) or other vehicles.

For all these cases where the class is very similar to our target class (see Fig. 13), we adopt a similar strategy as in (Mathias et al. 2014). That is, a method is neither penalized nor rewarded for tracking or not tracking those similar classes. Since a detector is likely to fire in those cases, we do not want to penalize a tracker with a set of false positives for properly following that set of detections, i.e., of a person on a bicycle. Likewise, we do not want to penalize with false negatives a tracker that is based on motion cues and therefore does not track a sitting person.

To handle these special cases, we adapt the tracker-to-target assignment algorithm to perform the following steps:
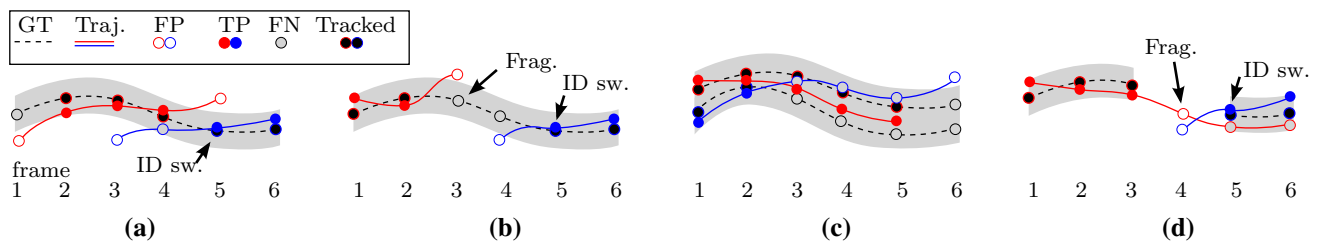
Fig. 12 Four cases illustrating tracker-to-target assignments. **a** An ID switch occurs when the mapping switches from the previously assigned red track to the blue one. **b** A track fragmentation is counted in frame 3 because the target is tracked in frames 1–2, then interrupts, and then reacquires its 'tracked' status at a later point. A new (blue) track hypothesis also causes an ID switch at this point. **c** Although the tracking results are reasonably good an optimal single-frame assignment in frame 1 is propagated through the sequence, causing 5 missed targets (FN) and 4 false positives (FP). Note that no fragmentations are counted in frames 3 and 6 because tracking of those targets is not resumed at a later point. **d** A degenerate case illustrating that target re-identification is not handled correctly. An interrupted ground-truth trajectory will typically cause a fragmentation. Also note the less intuitive ID switch, which is counted because blue is the closest target in frame 5 that is not in conflict with the mapping in frame 4

Fig. 13 The annotations include different classes of objects similar to the target class, a pedestrian in our case. We consider these special classes (distractor, reflection, static person and person on vehicle) to be so similar to the target class that a tracker should neither be penalized nor rewarded for tracking them in the sequence (Color figure online)

1. At each frame, all bounding boxes of the result file are matched to the ground truth via the Hungarian algorithm.
2. All result boxes that overlap more than the matching threshold ($> 50\%$) with one of these classes (distractor, static person, reflection, person on vehicle) excluded from the evaluation.
3. During the final evaluation, *only* those boxes that are annotated as *pedestrians* are used.

References

Alahi, A., Ramanathan, V., & Fei-Fei, L. (2014). Socially-aware large-scale crowd forecasting. In *Conference on computer vision and pattern recognition*.

Andriluka, M., Roth, S., & Schiele, B. (2010). Monocular 3D pose estimation and tracking by detection. In *Conference on computer vision and pattern recognition*.

Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., & Schiele, B. (2018). Posetrack: A benchmark for human pose estimation and tracking. In *Conference on computer vision and pattern recognition*.

Babaee, M., Li, Z., & Rigoll, G. (2019). A dual CNN-RNN for multiple people tracking. *Neurocomputing*, *368*, 69–83.

Bae, S.-H., & Yoon, K.-J. (2014). Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Conference on computer vision and pattern recognition*.

Bae, S.-H., & Yoon, K.-J. (2018). Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *Transactions on Pattern Analysis and Machine Intelligence*, *40*(3), 595–610.

Baisa, N. L. (2018). Online multi-target visual tracking using a HISP filter. In *International joint conference on computer vision, imaging and computer graphics theory and applications*.

Baisa, N. L. (2019a). Online multi-object visual tracking using a GM-PHD filter with deep appearance learning. In *International conference on information fusion*.

Baisa, N. L. (2019b). Occlusion-robust online multi-object visual tracking using a GM-PHD filter with a CNN-based re-identification. arXiv preprint arXiv:1912.05949.

Baisa, N. L. (2019c). Robust online multi-target visual tracking using a HISP filter with discriminative deep appearance learning. arXiv preprint arXiv:1908.03945.

Baisa, N. L., & Wallace, A. (2019). Development of a n-type GM-PHD filter for multiple target, multiple type visual tracking. *Journal of Visual Communication and Image Representation*, *59*, 257–271.

Baker, S., Scharstein, D., Lewis, J. P., Roth, S., Black, M. J., & Szeliski, R. (2011). A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, *92*(1), 1–31.

Ban, Y., Ba, S., Alameda-Pineda, X., & Horaud, R. (2016). Tracking multiple persons based on a variational Bayesian model. In *European conference on computer vision workshops*.

Battaglia, P., Pascanu, R., Lai, M., Rezende, D. J., et al. (2016). Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*.

Benfold, B., & Reid, I. (2011). Unsupervised learning of a scene-specific coarse gaze estimator. In *International conference on computer vision*.

Bergmann, P., Meinhardt, T., & Leal-Taixé, L. (2019). Tracking without bells and whistles. In *International conference on computer vision*.

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Image and Video Processing,*. https://doi.org/10.1155/2008/246309.

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016a). Simple online and realtime tracking. In *International conference on image processing*.

Bewley, A., Ott, L., Ramos, F., & Upcroft, B. (2016b). Alextrac: Affinity learning by exploring temporal reinforcement within association chains. In *International conference on robotics and automation*.

Bochinski, E., Eiselein, V., & Sikora, T. (2017). High-speed tracking-by-detection without using image information. In *International conference on advanced video and signal based surveillance*.

Boragule, A., & Jeon, M. (2017). Joint cost minimization for multi-object tracking. *International conference on advanced video and signal based surveillance*.

Brasó, G., & Leal-Taixé, L. (2020). Learning a neural solver for multiple object tracking. In *Conference on computer vision and pattern recognition*.

Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., & Hays, J. (2019). Argoverse: 3D tracking and forecasting with rich maps. In *Conference on computer vision and pattern recognition*.

Chen, J., Sheng, H., Zhang, Y., & Xiong, Z. (2017a). Enhancing detection model for multiple hypothesis tracking. In *Conference on computer vision and pattern recognition workshops*.

Chen, L., Ai, H., Chen, R., & Zhuang, Z. (2019). Aggregate tracklet appearance features for multi-object tracking. *Signal Processing Letters*, *26*(11), 1613–1617.

Chen, W., Chen, X., Zhang, J., & Huang, K. (2017b). Beyond triplet loss: A deep quadruplet network for person re-identification. In *Conference on computer vision and pattern recognition*.

Choi, W. (2015). Near-online multi-target tracking with aggregated local flow descriptor. In *International conference on computer vision*.

Chu, P., Fan, H., Tan, C. C., & Ling, H. (2019). Online multi-object tracking with instance-aware tracker and dynamic model refreshment. In *Winter conference on applications of computer vision*.

Chu, P., & Ling, H. (2019). FAMNet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *International conference on computer vision*.

Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., & Yu, N. (2017). Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism. In *International conference on computer vision*.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Conference on computer vision and pattern recognition workshops*.

Dave, A., Khurana, T., Tokmakov, P., Schmid, C., & Ramanan, D. (2020) Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision*.

Dehghan, A., Assari, S. M., & Shah, M. (2015) GMMCP-tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Conference on computer vision and pattern recognition workshops*.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., & Leal-Taixe, L. (2019). Cvpr19 tracking and detection challenge: How crowded can it get? arXiv preprint arXiv:1906.04567.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., & Leal-Taixé, L. (2020). MOT20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003.

Dicle, C., Camps, O., & Sznaier, M. (2013) The way they move: Tracking targets with similar appearance. In *International conference on computer vision*.

Dollár, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast feature pyramids for object detection. *Transactions on Pattern Analysis and Machine Intelligence*, *36*(8), 1532–1545.

Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2009) Pedestrian detection: A benchmark. In *Conference on computer vision and pattern recognition workshops*.

Eiselein, V., Arp, D., Pätzold, M., & Sikora, T. (2012). Real-time multi-human tracking using a probability hypothesis density filter and multiple detectors. In *International conference on advanced video and signal-based surveillance*.

Ess, A., Leibe, B., Schindler, K., & Van Gool, L. (2008). A mobile vision system for robust multi-person tracking. In *Conference on computer vision and pattern recognition*.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, *111*(1), 98–136.

Fagot-Bouquet, L., Audigier, R., Dhome, Y., & Lerasle, F. (2015). Online multi-person tracking based on global sparse collaborative representations. In *International conference on image processing*.

Fagot-Bouquet, L., Audigier, R., Dhome, Y., & Lerasle, F. (2016). Improving multi-frame data association with sparse representations for robust near-online multi-object tracking. In *European conference on computer vision workshops*.

Fang, K., Xiang, Y., Li, X., & Savarese, S. (2018). Recurrent autoregressive networks for online multi-object tracking. In *Winter conference on applications of computer vision*.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2006) Efficient belief propagation for early vision. In *Conference on computer vision and pattern recognition*.

Ferryman, J., & Ellis, A. (2010) PETS2010: Dataset and challenge. In *International conference on advanced video and signal based surveillance*.

Ferryman, J., & Shahrokni, A. (2009). PETS2009: Dataset and challenge. In *International workshop on performance evaluation of tracking and surveillance*.

Fu, Z., Angelini, F., Chambers, J., & Naqvi, S. M. (2019). Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking. *Transactions on Multimedia*, *21*(9), 2277–2291.

Fu, Z., Feng, P., Angelini, F., Chambers, J. A., & Naqvi, S. M. (2018). Particle PHD filter based multiple human tracking using online group-structured dictionary learning. *Access*, *6*, 14764–14778.

Geiger, A., Lauer, M., Wojek, C., Stiller, C., & Urtasun, R. (2014). 3D traffic scene understanding from movable platforms. *Transactions on Pattern Analysis and Machine Intelligence*, *36*(5), 1012–1025.

Geiger, A., Lenz, P., & Urtasun, R. (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Conference on computer vision and pattern recognition*.

Girshick, R. (2015). Fast R-CNN. In *International conference on computer vision*.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Conference on computer vision and pattern recognition*.

Held, D., Thrun, S., & Savarese, S. (2016). Learning to track at 100 fps with deep regression networks. In *European conference on computer vision*.

Henriques, J. a., Caseiro, R., & Batista, J. (2011). Globally optimal solution to multi-object tracking with merged measurements. In *International conference on computer vision*.

Henschel, R., Leal-Taixé, L., Cremers, D., & Rosenhahn, B. (2018). Fusion of head and full-body detectors for multi-object tracking. In *Conference on computer vision and pattern recognition workshops*.

Henschel, R., Zou, Y., & Rosenhahn, B. (2019). Multiple people tracking using body and joint detections. In *Conference on computer vision and pattern recognition workshops*.

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

Ju, J., Kim, D., Ku, B., Han, D., & Ko, H. (2017a). Online multi-object tracking with efficient track drift and fragmentation handling. *Journal of the Optical Society of America A*, *34*(2), 280–293.

Ju, J., Kim, D., Ku, B., Han, D. K., & Ko, H. (2017b). Online multi-person tracking with two-stage data association and online appearance model learning. *IET Computer Vision*, *11*(1), 87–95.

Karunasekera, H., Wang, H., & Zhang, H. (2019). Multiple object tracking with attention to appearance, structure, motion and size. *Access*,. https://doi.org/10.1109/ACCESS.2019.2932301.

Kesten, R., Usman, M., Houston, J., Pandya, T., Nadhamuni, K., et al. (2019) Lyft level 5 av dataset 2019. https://level5.lyft.com/dataset/.

Keuper, M., Tang, S., Andres, B., Brox, T., & Schiele, B. (2018). Motion segmentation and multiple object tracking by correlation co-clustering. *Transactions on Pattern Analysis and Machine Intelligence,*. https://doi.org/10.1109/TPAMI.2018.2876253.

Kieritz, H., Becker, S., Häbner, W., & Arens, M. (2016). Online multi-person tracking using integral channel features. In *International conference on advanced video and signal based surveillance*.

Kim, C., Li, F., Ciptadi, A., & Rehg, J. M. (2015). Multiple hypothesis tracking revisited. In *International conference on computer vision*.

Kim, C., Li, F., & Rehg, J. M. (2018). Multi-object tracking with neural gating using bilinear LSTM. In *European conference on computer vision*.

Kristan, M., et al. (2014). The visual object tracking VOT2014 challenge results. In *European conference on computer vision workshops*.

Kuhn, H. W., & Yaw, B. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2, 83–97.

Kutschbach, T., Bochinski, E., Eiselein, V., & Sikora, T. (2017). Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data. In *International conference on advanced video and signal based surveillance*.

Lan, L., Wang, X., Zhang, S., Tao, D., Gao, W., & Huang, T. S. (2018). Interacting tracklets for multi-object tracking. *Transactions on Image Processing*, 27(9), 4585–4597.

Le, N., Heili, A., & Odobez, J.-M. (2016). Long-term time-sensitive costs for CRF-based tracking by detection. In *European conference on computer vision workshops*.

Leal-Taixe, L., Canton-Ferrer, C., & Schindler, K. (2016). Learning by tracking: Siamese CNN for robust target association. In *Conference on computer vision and pattern recognition workshops*.

Leal-Taixé, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., & Savarese, S. (2014). Learning an image-based motion context for multiple people tracking. In *Conference on computer vision and pattern recognition*.

Leal-Taixé, L., Pons-Moll, G., & Rosenhahn, B. (2011). Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *International conference on computer vision workshops*.

Lee, S., & Kim, E. (2019). Multiple object tracking via feature pyramid Siamese networks. *Access*, 7, 8181–8194.

Lee, S.-H., Kim, M.-Y., & Bae, S.-H. (2018). Learning discriminative appearance models for online multi-object tracking with appearance discriminability measures. *Access*, 6, 67316–67328.

Levinkov, E., Uhrig, J., Tang, S., Omran, M., Insafutdinov, E., Kirillov, A., Rother, C., Brox, T., Schiele, B., & Andres, B. (2017). Joint graph decomposition and node labeling: Problem, algorithms, applications. In *Conference on computer vision and pattern recognition*.

Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High performance visual tracking with Siamese region proposal network. In *Conference on computer vision and pattern recognition*.

Li, Y., Huang, C., & Nevatia, R. (2009). Learning to associate: Hybrid boosted multi-target tracker for crowded scene. In *Conference on computer vision and pattern recognition*.

Liu, Q., Liu, B., Wu, Y., Li, W., & Yu, N. (2019). Real-time online multi-object tracking in compressed domain. *Access*, 7, 76489–76499.

Long, C., Haizhou, A., Chong, S., Zijie, Z., & Bo, B. (2017). Online multi-object tracking with convolutional neural networks. In *International conference on image processing*.

Long, C., Haizhou, A., Zijie, Z., & Chong, S. (2018) Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *International conference on multimedia and expo*.

Loumponias, K., Dimou, A., Vretos, N., & Daras, P. (2018). Adaptive tobit Kalman-based tracking. In *International conference on signal-image technology & internet-based systems*.

Ma, C., Yang, C., Yang, F., Zhuang, Y., Zhang, Z., Jia, H., & Xie, X. (2018a). Trajectory factory: Tracklet cleaving and re-connection by deep Siamese bi-GRU for multiple object tracking. In *International conference on multimedia and expo*.

Ma, L., Tang, S., Black, M. J., & Van Gool, L. (2018b). Customized multi-person tracker. In *Asian conference on computer vision*.

Mahgoub, H., Mostafa, K., Wassif, K. T., & Farag, I. (2017). Multi-target tracking using hierarchical convolutional features and motion cues. *International Journal of Advanced Computer Science & Applications*, 8(11), 217–222.

Maksai, A., & Fua, P. (2019). Eliminating exposure bias and metric mismatch in multiple object tracking. In *Conference on computer vision and pattern recognition*.

Manen, S., Timofte, R., Dai, D., & Gool, L. V. (2016). Leveraging single for multi-target tracking using a novel trajectory overlap affinity measure. In *Winter conference on applications of computer vision*.

Mathias, M., Benenson, R., Pedersoli, M., & Gool, L. V. (2014). Face detection without bells and whistles. In *European conference on computer vision workshops*.

McLaughlin, N., Martinez Del Rincon, J., Miller, P. (2015). Enhancing linear programming with motion modeling for multi-target tracking. In *Winter conference on applications of computer vision*.

Milan, A., Leal-Taixé, L., Schindler, K., & Reid, I. (2015). Joint tracking and segmentation of multiple targets. In *Conference on computer vision and pattern recognition*.

Milan, A., Rezatofighi, S. H., Dick, A., Reid, I., & Schindler, K. (2017). Online multi-target tracking using recurrent neural networks. In *Conference on artificial on intelligence*.

Milan, A., Roth, S., & Schindler, K. (2014). Continuous energy minimization for multitarget tracking. *Transactions on Pattern Analysis and Machine Intelligence*, 36(1), 58–72.

Milan, A., Schindler, K., & Roth, S. (2013). Challenges of ground truth evaluation of multi-target tracking. In *Conference on computer vision and pattern recognition workshops*.

Milan, A., Schindler, K., & Roth, S. (2016). Multi-target tracking by discrete-continuous energy minimization. *Transactions on Pattern Analysis and Machine Intelligence*, 38(10), 2054–2068.

Nguyen Thi Lan Anh, F. N., Khan, Furqan, & Bremond, F. (2017). Multi-object tracking using multi-channel part appearance representation. In *International conference on advanced video and signal based surveillance*.

Pedersen, M., Haurum, J. B., Bengtson, S. H., & Moeslund, T. B. (June 2020). 3D-ZEF: A 3D zebrafish tracking benchmark dataset. In *Conference on computer vision and pattern recognition*.

Pirsiavash, H., Ramanan, D., & Fowlkes, C. C. (2011). Globally-optimal greedy algorithms for tracking a variable number of objects. In *Conference on computer vision and pattern recognition*.

Reid, D. B. (1979). An algorithm for tracking multiple targets. *Transactions on Automatic Control*, 24(6), 843–854.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*.

Rezatofighi, H., Milan, A., Zhang, Z., Shi, Q., Dick, A., & Reid, I. (2015). Joint probabilistic data association revisited. In *International conference on computer vision*.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

Sadeghian, A., Alahi, A., Savarese, S. (2017). Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *International conference on computer vision*.

Sanchez-Matilla, R., Cavallaro, A. (2019). A predictor of moving objects for first-person vision. In *International conference on image processing*.

Sanchez-Matilla, R., Poiesi, F., & Cavallaro, A. (2016). Online multi-target tracking with strong and weak detections. In *European conference on computer vision workshops*.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, *47*(1), 7–42.

Schuhmacher, D., Vo, B.-T., & Vo, B.-N. (2008). A consistent metric for performance evaluation of multi-object filters. *Transactions on Signal Processing*, *56*(8), 3447–3457.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., & Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Conference on computer vision and pattern recognition*.

Sheng, H., Chen, J., Zhang, Y., Ke, W., Xiong, Z., & Yu, J. (2018a). Iterative multiple hypothesis tracking with tracklet-level association. *Transactions on Circuits and Systems for Video Technology*, *29*(12), 3660–3672.

Sheng, H., Hao, L., Chen, J., et al. (2017). Robust local effective matching model for multi-target tracking. In *Advances in multimedia information processing* (Vol. 127, No. 8).

Sheng, H., Zhang, X., Zhang, Y., Wu, Y., & Chen, J. (2018b). Enhanced association with supervoxels in multiple hypothesis tracking. *Access*, *7*, 2107–2117.

Sheng, H., Zhang, Y., Chen, J., Xiong, Z., & Zhang, J. (2018c). Heterogeneous association graph fusion for target association in multiple object tracking. *Transactions on Circuits and Systems for Video Technology*, *29*(11), 3269–3280.

Shi, X., Ling, H., Pang, Y. Y., Hu, W., Chu, P., & Xing, J. (2018). Rank-1 tensor approximation for high-order association in multi-target tracking. *International Journal of Computer Vision*, *127*, 1063–1083.

Smith, K., Gatica-Perez, D., Odobez, J.-M., & Ba, S. (2005). Evaluating multi-object tracking. In *Workshop on empirical evaluation methods in computer vision*.

Son, J., Baek, M., Cho, M., & Han, B. (2017). Multi-object tracking with quadruplet convolutional neural networks. In *Conference on computer vision and pattern recognition*.

Song, Y., & Jeon, M. (2016). Online multiple object tracking with the hierarchically adopted GM-PHD filter using motion and appearance. In *International conference on consumer electronics*.

Song, Y., Yoon, Y., Yoon, K., & Jeon, M. (2018). Online and real-time tracking with the GMPHD filter using group management and relative motion analysis. In *International conference on advanced video and signal based surveillance*.

Song, Y., Yoon, K., Yoon, Y., Yow, K., & Jeon, M. (2019). Online multi-object tracking with GMPHD filter and occlusion group management. *Access*, *7*, 165103–165121.

Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J. S., Mostefa, D., & Soundararajan, P. (2006). The clear 2006 evaluation. In *Multimodal technologies for perception of humans*.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., & Caine, B., et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Conference on computer vision and pattern recognition*.

Tang, S., Andres, B., Andriluka, M., & Schiele, B. (2015). Subgraph decomposition for multi-target tracking. In *Conference on computer vision and pattern recognition*.

Tang, S., Andres, B., Andriluka, M., & Schiele, B. (2016). Multi-person tracking by multicuts and deep matching. In *European conference on computer vision workshops*.

Tang, S., Andriluka, M., Andres, B., & Schiele, B. (2017). Multiple people tracking with lifted multicut and person re-identification. In *Conference on computer vision and pattern recognition*.

Tao, Y., Chen, J., Fang, Y., Masaki, I., & Horn, B. K. (2018). Adaptive spatio-temporal model based multiple object tracking in video sequences considering a moving camera. In *International conference on universal village*.

Taskar, B., Guestrin, C., & Koller, D. (2003). Max-margin Markov networks. In *Advances in neural information processing systems*.

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics (intelligent robotics and autonomous agents)*. Cambridge: The MIT Press.

Tian, W., Lauer, M., & Chen, L. (2019). Online multi-object tracking using joint domain information in traffic scenarios. *Transactions on Intelligent Transportation Systems*, *21*(1), 374–384.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Conference on computer vision and pattern recognition*.

Wang, B., Wang, L., Shuai, B., Zuo, Z., Liu, T., et al. (2016). Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association. In *Conference on computer vision and pattern recognition*.

Wang, G., Wang, Y., Zhang, H., Gu, R., & Hwang, J.-N. (2019). Exploit the connectivity: Multi-object tracking with trackletnet. In *International conference on multimedia*.

Wang, S., & Fowlkes, C. (2016). Learning optimal parameters for multi-target tracking with contextual interactions. *International Journal of Computer Vision*, *122*(3), 484–501.

Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M., Qi, H., et al. (2020). UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, *193*, 102907.

Wen, L., Li, W., Yan, J., Lei, Z., Yi, D., & Li, S. Z. (2014). Multiple target tracking based on undirected hierarchical relation hypergraph. In *Conference on computer vision and pattern recognition*.

Wojke, N., & Paulus, D. (2016). Global data association for the probability hypothesis density filter using network flows. *International conference on robotics and automation*.

Wu, B., & Nevatia, R. (2006). Tracking of multiple, partially occluded humans based on static body part detection. In *Conference on computer vision and pattern recognition*.

Wu, H., Hu, Y., Wang, K., Li, H., Nie, L., & Cheng, H. (2019). Instance-aware representation learning and association for online multi-person tracking. *Pattern Recognition*, *94*, 25–34.

Xiang, J., Xu, G., Ma, C., & Hou, J. (2020). End-to-end learning deep CRF models for multi-object tracking. *Transactions on Circuits and Systems for Video Technology*, https://doi.org/10.1109/TCSVT.2020.2975842.

Xiang, Y., Alahi, A., & Savarese, S. (2015). Learning to track: Online multi-object tracking by decision making. In *International conference on computer vision*.

Xu, J., Cao, Y., Zhang, Z., & Hu, H. (2019). Spatial-temporal relation networks for multi-object tracking. In *International conference on computer vision*.

Xu, Y., Osep, A., Ban, Y., Horaud, R., Leal-Taixe, L., & Alameda-Pineda, X. (2020). How to train your deep multi-object tracker. In *Conference on computer vision and pattern recognition*.

Yang, F., Choi, W., & Lin, Y. (2016). Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In *Conference on computer vision and pattern recognition*.

Yang, M., & Jia, Y. (2016). Temporal dynamic appearance modeling for online multi-person tracking. *Computer Vision and Image Understanding*, https://doi.org/10.1016/j.cviu.2016.05.003.

Yang, M., Wu, Y., & Jia, Y. (2017). A hybrid data association framework for robust online multi-object tracking. *Transactions on Image Processing*, https://doi.org/10.1109/TIP.2017.2745103.

Yoon, J., Yang, H., Lim, J., & Yoon, K. (2015). Bayesian multi-object tracking using motion context from multiple objects. In *Winter conference on applications of computer vision*.

Yoon, J. H., Lee, C. R., Yang, M. H., & Yoon, K. J. (2016). Online multi-object tracking via structural constraint event aggregation. In *International conference on computer vision and pattern recognition*.

Yoon, K., Gwak, J., Song, Y., Yoon, Y., & Jeon, M. (2020). OneShotDa: Online multi-object tracker with one-shot-learning-based data association. *Access*, *8*, 38060–38072.

Yoon, K., Kim, D. Y., Yoon, Y.-C., & Jeon, M. (2019a). Data association for multi-object tracking via deep neural networks. *Sensors*, *19*, 559.

Yoon, Y., Boragule, A., Song, Y., Yoon, K., & Jeon, M. (2018a). Online multi-object tracking with historical appearance matching and scene adaptive detection filtering. In *International conference on advanced video and signal based surveillance*.

Yoon, Y., Kim, D. Y., Yoon, K., Song, Y., & Jeon, M. (2019b). Online multiple pedestrian tracking using deep temporal appearance matching association. arXiv preprint arXiv:1907.00831.

Yoon, Y.-C., Song, Y.-M., Yoon, K., & Jeon, M. (2018). Online multi-object tracking using selective deep appearance matching. In *International conference on consumer electronics Asia*.

Zamir, A. R., Dehghan, A., & Shah, M. (2012). GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs. In *European conference on computer vision*.

Zhang, L., Li, Y., & Nevatia, R. (2008). Global data association for multi-object tracking using network flows. In *Conference on computer vision and pattern recognition*.

Zhang, Y., Sheng, H., Wu, Y., Wang, S., Lyu, W., Ke, W., et al. (2020). Long-term tracking with deep tracklet association. *Transactions on Image Processing*, *29*, 6694–6706.

Zhou, H., Ouyang, W., Cheng, J., Wang, X., & Li, H. (2018). Deep continuous conditional random fields with asymmetric inter-object constraints for online multi-object tracking. *Transactions on Circuits and Systems for Video Technology*,. https://doi.org/10.1109/TCSVT.2018.2825679.

Zhou, X., Jiang, P., Wei, Z., Dong, H., & Wang, F. (2018b). Online multi-object tracking with structural invariance constraint. In *British machine vision conference*.

Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., & Yang, M.-H. (2018). Online multi-object tracking with dual matching attention networks. In *European conference on computer vision workshops*.