Large scale Real-world Multi-Person Tracking

Bing Shuai, Alessandro Bergamo, Uta Buechler
Andrew Berneshawi, Alyssa Boden, Joseph Tighe

AWS AI Labs
https://github.com/amazon-research/tracking-dataset

Abstract. This paper presents a new large scale multi-person tracking dataset. Our dataset is over an order of magnitude larger than currently available high quality multi-object tracking datasets such as MOT17, HiEve, and MOT20 datasets. The lack of large scale training and test data for this task has limited the community's ability to understand the performance of their tracking systems on a wide range of scenarios and conditions such as variations in person density, actions being performed, weather, and time of day. Our dataset was specifically sourced to provide a wide variety of these conditions and our annotations include rich meta-data such that the performance of a tracker can be evaluated along these different dimensions. The lack of training data has also limited the ability to perform end-to-end training of tracking systems. As such, the highest performing tracking systems all rely on strong detectors trained on external image datasets. We hope that the release of this dataset will enable new lines of research that take advantage of large scale video based training data.

Keywords: Multi-object tracking, dataset, MOT

1 Introduction

Large-scale datasets are the fuel that has driven the success of learning-based methods over the past decade. The introduction of large datasets, such as ImageNet[21], MSCOCO[34], LSUN[63] and Kinetics[11], has enabled the development of deep learning-based models which have rapidly advanced the field of computer vision. Unfortunately, no such large scale dataset has been collected for multi-object tracking to date. The multi-object tracking task [8, 57, 66, 50, 65, 41] requires detection and ID assignment of all objects for each frame in a video. In practice many current datasets have people as the only objects (multi-person), which will also be our focus. The most popular datasets used today, MOT17 [39] and MOT20 [20], have just 14 and 8 videos respectively, greatly limiting the ability of researchers to develop data hungry models that require large tracking datasets as well as limiting the measure of generalizability of tracking methods given the small number of videos used for testing. In this work we present a new multi-person tracking dataset that is an order of magnitude larger than MOT17 [39] and MOT20 [20], while maintaining the high quality bar of annotation present in those datasets.

One reason for the lack of large scale multi-object tracking datasets is the significant cost to collecting such a dataset. The collection and annotation of these datasets is non-trivial as both the curation (sourcing) and labeling require significantly higher manual human labor than classification or detection based datasets. For person tracking, sourcing video is particularly challenging because though there is a large volume of video content available on the internet, it is mostly content that does not align with our target video domain or the content rights are restricted such that the videos cannot be easily included in an academic dataset. The Kinetics[11] dataset, for example, attempted to remove this challenge by only providing links to YouTube videos but over time those videos were removed, leaving researchers with incomplete train and test sets and making it difficult to reliably compare to other works.

In this work we collect videos from sources where we are given the rights to redistribute the content and participants have given explicit consent, such as the MEVA[17] dataset. Our dataset consists of 236 videos captured mostly from static-mounted cameras. Approximately 80% of these videos are carefully sourced from scratch from stock footage websites and 20% are collected from existing datasets such as PathTrack[37] or MEVA[17]. While building the dataset we place special importance on sourcing indoor and outdoor videos with different lightning conditions, diverse camera angles (from birds-eye view to low-angle view), varying weather conditions (sunny, raining, cloudy, night), various levels of occlusion and different crowd densities. Section 3 presents a detailed analysis of these factors.

In addition to sourcing, collecting high quality annotations is especially challenging for multi-object tracking datasets. This is largely due to the complexity of the task. Classification datasets [21, 11] only require one or more labels to be tagged per entire image or video whereas detection datasets [34, 24] increase the complexity by not only requiring a list of objects, but also the object's location specified by a bounding box. Multi-object tracking extends the idea of object detection even further by also requiring a unique object identifier for every labeled bounding box throughout a video recording. This annotation task is especially challenging in crowded scenes where even a human annotator could easily lose or confuse an object with another if they get partly or fully occluded.

In this work we adopt a two stage annotation pipeline that leverages AWS SageMaker GroundTruth (an iteration of Amazon MTurk). When annotating videos for tracking, many edge cases emerge and must be handled consistently to have a meaningful measure of an algorithm's quality. In our annotation process, we have thoroughly considered edge cases such as people with high occlusion or person reflections and defined strict protocols for dealing with each edge case. For example, we annotate reflections of people but tag such annotations specifically so they can be properly handled during training and evaluation. After carefully defining our annotation criteria, we use our trained workforce to annotate all videos from scratch. More details regarding our annotation protocol can be found in Section 4.

We demonstrate the benefit our large-scale dataset adds to the community by (1) comparing key statistics with existing MOT benchmarks (Section 5) and (2) training and evaluating state-of-the-art multi-object tracking models on our dataset (Section 6). The latter shows that our benchmark contains many challenging scenarios where current state-of-the-art models fail to perform well. We hope that the publication of our dataset will drive the tracking community towards developing more robust models that can generalize to a wide variety of smart home/city scenarios.

2 Related Work

Multi-Object Tracking Datasets. MOT is an essential part of important applications such as autonomous driving [23, 43, 45], smart city [18, 38, 12, 13] and activity recognition [58, 5]. Especially the field of autonomous driving has grown significantly, which is also reflected in the number of large-scale benchmarks published for this scenario [27, 16, 31, 52, 9, 14, 22, 62]. Some of these datasets have also been used to train and/or evaluate person tracking models [42, 50]. The challenges such benchmarks entail are fast camera motions and quick position changes of pedestrians. However, the amount of occlusions and crowdness is rather limited and thus not sufficient enough to train robust tracking models that can operate in high-occlusion scenarios. In contrast, synthetic datasets that have been specifically created for pedestrian detection/tracking in urban scenarios [25, 26] contain scenes with varying person densities and can therefore be very valuable for person tracking. The clear advantage is that they do not require any manual annotations. Although the quality of synthetic data improves steadily, the usage of such data is rather limited due to the apparent domain shift to real-world data.

Recently, a few real-world MOT datasets have been proposed. For instance, CroHD [53] dataset was introduced to track pedestrain's head in crowded scenes, GMOT-40 [4] was proposed for the purpose of general object tracking, and MVMHAT [61] and MMPTRACK [29] are adopted for multi-camera multi-person tracking. In general, their sizes are a magnitude smaller than our dataset. One of the biggest real-world MOT datasets is TAO [19], which provides a great variety of scenes. Since TAO is created for general object tracking, the number of challenging person tracking sequences is rather limited given that a large number of videos contain only a single person. Moreover, TAO provides full annotations for only a small fraction of videos, which makes it difficult to train on. In contrast, our dataset has been exhaustively annotated.

Finally, the MOT datasets that are most comparable to ours are HiEve [35], MOT17 [39] and MOT20 [20]. HiEve consists of 32 videos (13.5% of the size of our dataset) and provides annotations for different human-centric understanding tasks such as pedestrian tracking or pose estimation [51, 36, 15]. The main goal of HiEve is to provide a set of videos that are recorded during complex events (e.g. earthquake escapes). Our dataset, on the other hand, has the objective of providing a wide variety of smart home/city scenarios during different seasons,

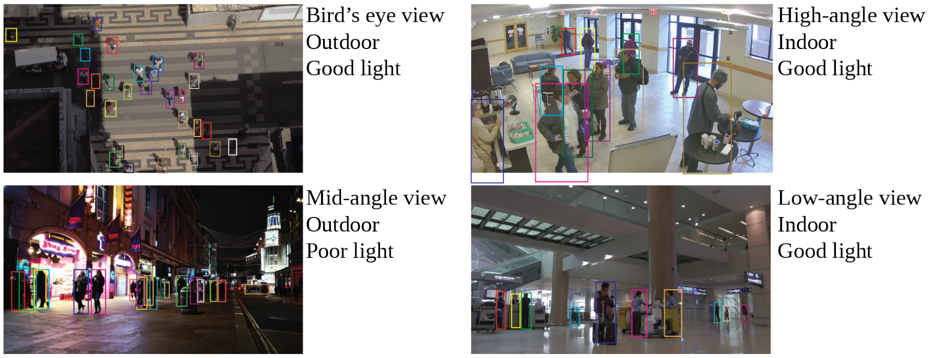


Fig. 1: The figure is best viewed in color. Frames in the video are exhaustively annotated with person boxes, each of which have a unique identifier (i.e. color-coded box). The videos in the dataset cover diverse tracking scenarios in terms of camera angles, weather / lighting condition and scenery types.

varying lightning and weather conditions and diverse crowd densities without focusing solely on the complexity of events. The most popular MOT benchmark which has also a similar purpose as ours is MOT17 [39]. The benchmark consists of 14 videos that are recorded at 9 different scenes with different lightning and camera angles. MOT20 [39] extends the MOT17 benchmark by 8 additional videos, which was specifically created for tracking in crowds. Our dataset also contains very crowded scenes, but provides on top of that a wide variety of pedestrian densities indoor and outdoor.

Multi-Object Tracking Methods. Many of the well-known MOT models follow the detection-by-tracking paradigm [50, 8, 57, 32, 47, 55, 59, 60], in which object instances are firstly detected for every frame and then they are linked across frames to form object tracks. Recently, online trackers [50, 67, 65, 42, 56] have steadily gained ground by pushing the results on MOTChallenge [39] to new highs. Those trackers are usually deep neural networks that include key models for online tracking, which include a detection model [67, 44, 54, 10], a motion model [8, 57, 50, 7, 33] and an optional person re-identification model [56, 65]. Those models are usually jointly trained with tracking annotations, i.e. a bounding box with a unique identifier. Due to the scarcity of those annotations, self-supervised training techniques [66, 50, 65, 49] were developed to leverage image-based object detection datasets for model pre-training. In this work, we train and evaluate three recent state-of-the-art online trackers on our dataset.

3 Video Sourcing

The creation of a dataset for training / evaluating person tracking algorithms needs to strike a balance between the need of (1) having videos that represent a large variety of real-world tracking applications; (2) having videos containing challenging scenarios for tracking algorithms (e.g. occlusions, small objects); and (3) ensuring that the data is collected in a responsible way such that it can be used in perpetuity. Following these guidelines, we source our dataset in two steps.

Data Source Selection. We select a pool of data sources based on the availability of video content suitable for tracking applications, as well as the presence of an appropriate license that allows the data to be used and remain available for academic research. We source videos from stock video services (Fillerstock [1], Pexels [2], Pixabay [3]) and from public academic datasets for human activity understanding (MEVA [17], Virat [40], PathTrack [37]) where proper licensing is available. The breakdown of the number of videos for each data source is provided in the supplementary material. Note that although MEVA and Virat come with incomplete person bounding boxes annotations, we re-annotate all videos included in our dataset to ensure consistency in annotations across all data sources. We first create an initial large candidate set of videos by automatically querying content from Fillerstock, Pexels, and Pixabay using a pre-defined set of search keywords such as "person walking in the shopping mall" (please refer to the supplementary material for the full list). The union of these videos and the videos from the public datasets form our candidate video set.

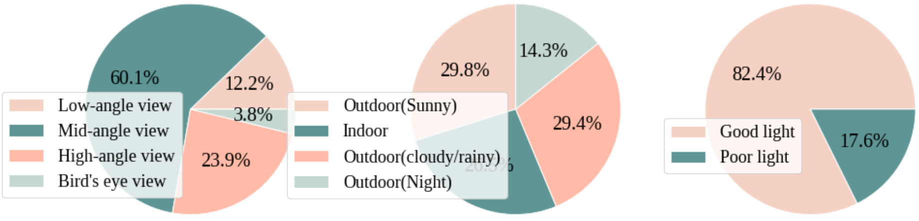
Manual Selection. Our initial candidate dataset includes 8000+ videos which are then manually inspected by a team of experts. The selection processes took into account the following criteria: (1) application aligned (fixed connected home or city level cameras), (2) moving crowds, (3) occlusion, (4) background variability, (5) static vs moving cameras, (6) camera position and (7) environment conditions (day/night, sunny/rain/snow/cloudy etc). More details to the mentioned criteria are elaborated in the supplementary material. In total, we select 236 videos for manual annotation and inclusion in our dataset. The cumulative temporal duration of these videos is 139 minutes.

4 Annotation Pipeline

Annotating person boxes with identities is time-consuming and error-prone. To this end, we adopt AWS SageMaker GroundTruth (SMGT) service[1] (an advanced version of Amazon Mechanical Turk). This workflow works as follows. First, the annotator draws bounding boxes for all visible people in the starting frame. In the next frame, the SMGT service leverages a pre-trained model to predict the bounding box for each annotated person. The annotator first verify the quality of predicted bounding boxes and adjust the bounding boxes as needed. Then, the annotator draws bounding boxes for those persons that do not appear in earlier frames.

We employ professional annotators that have been specially trained for this task. We ask them to annotate every possible visible person in the video unless they are too small in size ($< 20 \times 20$ pixels) to be accurately localized or they are

[1] `https://docs.aws.amazon.com/sagemaker/latest/dg/sms-video-object-tracking.html`



(a) Camera angles (b) Scenes / weathers (c) Light condition

Fig. 2: Video-level statistics of training videos in our dataset.

in a crowd. In the latter case, we ask the annotators to draw a bounding box with `crowd label` that includes all people in the crowd (e.g. Fig. 1(bottom left)). If a person enters the area labeled as crowd, with $> 95\%$ of the person's bounding box covered by a crowd box, we label this occurrence as 'ignore' to ensure that the predicted tracks are not penalized on these cases. As shown in Fig. 1, we annotate with `amodal` bounding boxes, indicating that the full extent of the bounding box is annotated regardless of the visibility status of the underlying person. In addition, we also annotate the corresponding `visible` bounding boxes that only enclose the visible part of the person body. This inclusion of both annotation types give researchers the most flexibility when choosing how to train their models and evaluating these models on other datasets.

To ensure that the annotation is of high quality, we perform a second round of labeling where a separate group of annotators checks if (1) all people are annotated, (2) all bounding boxes are correctly localized and (3) the identity of a person track is consistent throughout the video. In case the annotators notice a mistake, they correct the error. Finally, the authors of this paper do a final verification pass on the data, sending back any videos that have errors for re-annotation. This rigorous process allows us to have high confidence of the quality of the provided annotations. We first annotate at 5 frames per second. Then, we linearly interpolate those annotations and let our trained annotators verify the correctness of those interpolations for every frame and person.

Given that not all annotated person boxes are equally interesting, and some might even be perceived as noise, we further annotate each person track with the following tags: 1), sitting/standing still person; 2),person in vehicle; 3),person on open-vehicle; 4), reflection; 5), severely occluded person; 6), person in background; 7), foreground person. These fine-grained track-level tags enable to train or evaluate models along different sets of person tracks based on the needs of various tracking applications. The definition and visual examples of those tags are provided in the supplementary materials.

5 Dataset

To understand how our dataset compares to current MOT datasets we analyze various statistics of our and other publicly available datasets. We specifically compare to three popular datasets: MOT17 [39], HiEve [35], and MOT20 [20].

5.1 Video-level Statistics

Camera Angles. We categorize the angles of the cameras that are used to record the underlying videos into four buckets: (1) bird's-eye view, (2) high-angle view, (3) mid-angle view and (4) low-angle view. Visual examples are given in Fig. 1. As shown in Fig. 2a, our dataset contains 143 (60.1%) and 33 (23.9%) videos that are recorded by mid and low-angle-view cameras, respectively. On this front, the closest dataset to ours is MOT17 [39] that has 10 (71.4%) mid-angle-view and 4 (28.6%) high-angle-view videos. Out of 32 videos in HiEve dataset [35], only 2 (6.3%) videos are recorded with mid-angle-view cameras and the remaining 30 (93.7%) videos are with high-angle-view. For MOT20 dataset [20], all 8 videos are captured with high-angle-view cameras.

Scenes and Weather. We categorize the scene of a video into two buckets: (1) indoor (e.g. cafe house, mall, airport) and (2) outdoor (e.g. street, plaza, beach). For outdoor videos, we further annotate the weather condition. As shown in Fig. 2b, there are 63 (26.5%) indoor videos and the outdoor videos are evenly spread across three weather/light conditions (sunny, cloudy, night/dark). Furthermore as we show in Fig. 2c, there are 42 (17.6%) videos that have poor light condition, under which tracking people becomes increasingly challenging. Overall, our dataset provides a good diversity in terms of scene types and light conditions. In comparison, MOT17 [39] includes 2 indoor and 2 night videos.

People Density. We define the people density ($\texttt{d}$) of the scene to be the average number of people per frame, based on which we categorize each video into four buckets: low density ($\texttt{d} \leq 10$), medium density ($10 < \texttt{d} \leq 30$), high density ($30 < \texttt{d} \leq 60$) and extremely high density ($\texttt{d} > 60$). As shown in Fig. 3, our dataset has a similar distribution with MOT17 [39] and HiEve [35] dataset, although it has a significantly larger scale. Note that although there is a positive correlation between the tracking difficulty and the people density of the video when the camera angle and scene / light condition is similar, people density is not the only indicator of difficulty level of underlying videos. For example, tracking a person in a low-angle-view video with low density can be more challenging than that in a high-density bird's-eye video due to the high level of occlusion in the low-angle-view video.

In terms of the above factors, our dataset provides a set of videos that resembles a similar distribution with the current dataset MOT17 [39] but at an order of magnitude larger scale. Importantly our dataset is highly diverse which makes the training and evaluation of tracking models more representative to real-world person tracking challenges.

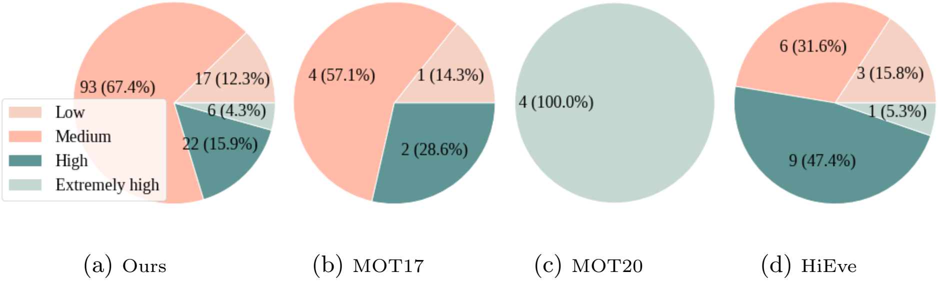


Fig. 3: Video-level people density distribution of training videos.

5.2 Track-level Statistics

We further analyze the statistics of each track annotated in our dataset. We represent a person track as a temporally ordered set of bounding boxes $\mathcal{T} = [\mathtt{bb}_{t_s}, \ldots, \mathtt{bb}_t, \ldots, \mathtt{bb}_{t_e}]$, in which t_s and t_e are the start and terminal timestamp of person track $\mathcal{T}$ respectively, $\mathtt{bb}_t = (x_t, y_t, w_t, h_t)$ where (x_t, y_t) is the center point coordinates of person bounding box at time t and w_t, h_t its width and height. In total, 12,150 unique person tracks are annotated, out of which 7,096 tracks are from training videos, and the remaining 5,054 from test videos. Furthermore, 7,534 tracks are labeled with "foreground person" tag, based on which we derive the statistics of person tracks as follows.

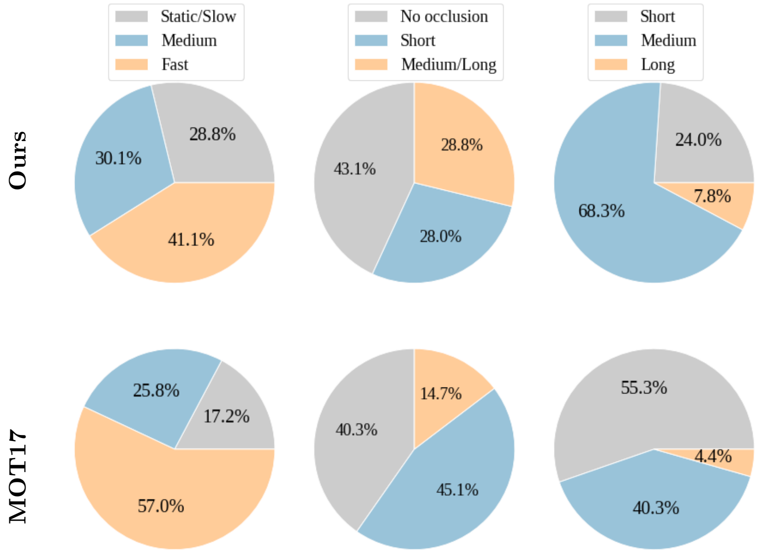
Average Track Speed. We define the temporally normalized motion vector $\mathbf{m}_{(t_1 \to t_2)}$ for person track $\mathcal{T}$ between timestamp t_1 and t_2 $(t_2 > t_1)$ as follows:

$$\mathbf{m}_{t_1 \to t_2} = \frac{1}{\zeta \cdot (t_2 - t_1)} (x_{t_2} - x_{t_1}, y_{t_2} - y_{t_1}) \tag{1}$$

in which ζ is the average length of the person bounding box at timestamp t_1 and t_2, that is $\zeta = 0.5 * (\sqrt{(w_{t_1} \cdot h_{t_1})} + \sqrt{(w_{t_2} \cdot h_{t_2})})$. Therefore, $\mathbf{m}_{(t_1 \to t_2)}$ indicates the direction of the person's motion between timestamp t_1 and t_2, and its $\mathbf{L}_2$ norm $||\mathbf{m}_{(t_1 \to t_2)}||_2$ reflects the speed of the corresponding person within a unit of time. Then, we derive the average speed $\mathbf{v}$ for a track with the following equation:

$$\mathbf{v}_{(\mathcal{T})} = \frac{1}{|\mathbf{T}| - 1} \sum_{i=2}^{|\mathbf{T}|} ||\mathbf{m}_{(\mathbf{T}[i-1] \to \mathbf{T}[i])}||_2 \tag{2}$$

where $\mathbf{T} = \{t_s, \ldots, t, \ldots, t_e\}$ is a sorted list of timestamps that the person appears. We bucketize each person track to have a static/slow, medium and fast speed if $\mathbf{v}_{(\mathcal{T})} < 0.2$, $0.2 \leq \mathbf{v}_{(\mathcal{T})} < 0.6$ and $\mathbf{v}_{(\mathcal{T})} \geq 0.6$. In Fig. 4a, we show the distribution of track speed of our dataset in comparison with MOT17 [39]. A few videos in MOT17 are recorded with moving camera, which leads to larger portion of higher-speed person tracks (e.g. a person standing still appears to be non-static in a video with moving background).



(a) Track speed (b) Occlusion duration (c) Track duration

Fig. 4: Key statistics of person tracks in training videos from our dataset and MOT17.

Occlusion Duration. A person becomes fully occluded if its appearance feature is not discernible at that particular time. In our case, this happens if the annotator is unable to locate their position without inferring from temporal context. Therefore, we define the occlusion duration of a person track to be the cumulative duration ($\mathbf{o}_{(\mathcal{T})}$) of the person being fully occluded. We further categorize each person track to have no, short and medium/long occlusion if $\mathbf{o}_{(\mathcal{T})} = 0$, $0 < \mathbf{o}_{(\mathcal{T})} < 2(s)$ and $\mathbf{o}_{(\mathcal{T})} \geq 2(s)$. As shown in Fig. 4b, our dataset includes a significantly higher portion of person tracks with medium/long occlusion in comparison to MOT17. In addition, there is a significant percentage (3.5%) of person tracks whose occlusion duration is longer than 10 seconds. A particular challenge in person / object tracking is to preserve the identity consistency before and after the object becomes fully occluded. In this respect, our dataset provides challenging and interesting cases.

Track Duration. The duration of a track ($\mathbf{l}_{(\mathcal{T})}$) is defined as the time range between the first and last appearance of the person in the video, that is $\mathbf{l}_{(\mathcal{T})} = t_e - t_s$. We classify each person track to be short, medium and long if $\mathbf{l}_{(\mathcal{T})} < 5(s)$, $5 \leq \mathbf{l}_{(\mathcal{T})} < 30(s)$ and $\mathbf{l}_{(\mathcal{T})} \geq 30(s)$ respectively. As shown in Fig. 4c, person tracks in our dataset tend to be longer in contrast to MOT17 [39]. Considering that tracking a person in a longer duration is both interesting and technically challenging, our dataset offers valuable testing cases along this aspect.

Dataset	#Videos	Length (secs)	#Annotated Frames	#Person Tracks	Min Res.	Min. FPS
HiEve[35]	19	1,842	32,929	1,736	352x258	15
MOT17[39]	7	215	5,316	546	640x480	14
MOT20[20]	4	357	8,931	2,215	1173x880	25
Ours	138	4,736	118,685	7,096	720x480	15

Table 1: Comparison of dataset statistics (of training set) between our and existing datasets. Annotated Frames refer to the frames that are manually annotated and those that are automatically interpolated and then manually verified.

In Tab. 1, we further compare our dataset with existing person tracking datasets. In comparison to MOT17 [39], the most popular dataset for multi-person tracking research, our dataset includes an order of magnitude larger number of unique person tracks and videos. Although MOT20 [20] includes more annotated person tracks, their scope is specifically for tracking people in crowds. Both the diversity of videos and the person tracks in our dataset are unparalleled w.r.t other dataset including HiEve [35], which makes it a more challenging and realistic evaluation benchmark for multi-person tracking.

5.3 Benchmarking

We randomly split the videos with 60% train and 40% test. To make sure that both subsets follow a similar distribution, we perform the split for each video source separately. Overall, there are 138 train and 98 test videos, and we treat it as the official split of this dataset. The statistics for both splits are listed in the supplementary materials.

We only evaluate on keyframes for bounding boxes with the "foreground person" and "standing / sitting still person" tag that aren't fully occluded. The key frames are identical with those used for manual annotation, so we are evaluating the results at 5FPS. With this evaluation protocol, we are discounting the influence from the detection failures but implicitly amplifying the effect from identity inconsistency. By doing this the missed detection on a fully occluded person are not penalized. We argue that it's more important to keep the identity prediction consistent before and after the person is fully occluded rather than inferring bounding boxes for a person that is not visible.

6 Experiments

We evaluate our dataset using three recent state-of-the-art online trackers, each including a person detection and person identity association model, which are jointly trained with tracking annotations. We briefly introduce the methods.

CenterTrack [66] is a single-stage online tracking model that performs joint detection and tracking and is built upon the CenterNet [67] framework. The model takes as input (1) the previous RGB frame, (2) the current RGB frame, and (3) a heatmap with the tracked object centers. The model predicts the object boxes for the current frame, conditioned on the tracking center points that are provided as input. In addition, the model outputs the estimated offset motion vectors, based on which an online solver is used to link the boxes across frames.

SiamMOT [50] is a two-stage tracking model which uses Faster R-CNN [44] for its person detection model. A Siamese-based tracker [30, 28] is incorporated in the network as a motion model to associate the detection bounding boxes across frames. In this work, we use the best-performing motion model, EMM, as suggested in the original paper.

FairMOT [65] is a single-stage tracking model that uses CenterNet [67] as person detection model. In addition to CenterNet, this method adopts a parallel branch to extract a feature vector (embedding) for each person instance. Finally, the affinity between the person's location and its embedding, together with a motion model (Kalman filter) are used to link detected people across frames.

We choose the above three models as they cover both single-stage and two-stage detection models. Besides, they cover two mainstream linking techniques: CenterTrack [66] and SiamMOT [50] use learned motion models for bounding box linking, whereas FairMOT [65] leverages the similarity of person embeddings.

Implementation details. All models use DLA-34 [64] as feature backbone, and they are pre-trained on the CrowdHuman dataset [48]. We use the official open-source implementations for all the algorithms provided by the original authors. We train and evaluate the model with `amodal` bounding boxes. Please refer to the supplementary materials for more details.

Evaluation metrics. Following other literature, we report standard tracking metrics including MOTA and IDF1. In general, MOTA measures the overall performance of the end-to-end tracking system by accounting for both the detection and data association performance. IDF1, on the other hand, specifically indicates the performance of predicted identity consistency. For more details on these metrics we refer the reader to [6, 46].

6.1 Model Evaluation

In Tab. 2, we show the results of three recent online trackers. In the default evaluation protocol [39, 20, 35], all valid person boxes on key frames are evaluated. Under such setting, all models achieve relatively low MOTA and IDF1 in comparison to the performance on MOT17 [39] and HiEve [35], which underscores the challenges of our dataset. As expected, the detection failure (False Positive (FP) and False Negative (FN)) heavily influences the MOTA metric. We observe that a significant number of detection failures results from missed detections when a person becomes fully occluded. As we elaborated in Sec. 5.3, we should not heavily penalize those missed detections as long as the predicted identity is consistent before and after the occlusion happens. To this end, we apply an occlusion filter process to exclude those boxes tagged as being fully

Methods	Occlusion Filter	IDF1 (↑)	MOTA (↑)	FP (↓)	FN (↓)	IDsw (↓)
CenterTrack [66]	✗	43.04	52.31	24611	107037	10487
SiamMOT [50]	✗	49.84	59.56	13268	98069	9201
FairMOT [65]	✗	56.52	54.29	14568	116495	5179
CenterTrack [66]	✓	46.36	59.28	24340	71550	10319
SiamMOT [50]	✓	53.71	67.52	13217	62543	8942
FairMOT [65]	✓	61.05	61.79	14540	80034	5095

Table 2: Result comparison on the test split of our dataset. Occlusion filter means that only bounding boxes without being tagged as occluded are used during evaluation.

occluded from evaluation. As shown in Tab. 2, FN is significantly decreased, which lifts MOTA by a large margin. Additionally, after applying the filter, a person track with occlusion is "reduced" in length, which in-turn benefits IDF1. Nonetheless, the improvement of IDF1 is less significant than that of MOTA.

As shown in Tab. 2, SiamMOT achieves significantly higher MOTA compared to CenterTrack and FairMOT. We conjecture that its underlying detector – Faster-RCNN [44] – works better than CenterNet [67] which underlies the other two tracking models. To validate it, we run inference of the two underlying detectors — FRCNN [44] and CenterNet [67] on the test set. FRCNN achieves 82.03% AP@0.5 and CenterNet achieves 78.51% AP@0.5.[2] Not surprisingly, FairMOT achieves a significantly better IDF1 than the other two motion-based tracking models, despite the fact that the detected boxes have more errors than that of SiamMOT. This result suggests that person re-identification is essential for tracking models to preserve the identity consistency of predicted tracks in the case of occlusion.

6.2 In-Depth Model Analysis

Small-Size Person Tracking. Being able to correctly track small scale objects is important for real-world application scenarios. We categorize a person track as *small* if the average areas of the associated bounding boxes is smaller than 0.5% relative to the video frame area. For example, any bounding box whose area is smaller than 50×90 for a standard 720p video is considered small in size. In our test set, 1,624 tracks are categorized as "small". As shown in Tab. 3a, there is a significant performance gap between tracking large-size and small-size persons on both MOTA and IDF1. This is expected as both detecting and re-identifying low-resolution objects remains a major challenge.

Static vs. Moving. In real-world scenarios, video sequences contain a mix of static and moving objects. For example, people might be sitting on chairs or benches (e.g. at a park or in a waiting room), as well as standing and not

[2] We encourage the researchers report detection AP@0.5 of their tracking models on our dataset.

Method	IDF1(↑)		MOTA(↑)	
	small	large	small	large
CenterTrack	34.3	52.7	34.5	70.1
SiamMOT	47.1	56.6	49.9	75.2
FairMOT	50.2	66.6	40.8	72.0

(a) Results for tracks associated to small-size vs large-size persons.

Method	IDF1(↑)		MOTA(↑)	
	static	moving	static	moving
CenterTrack	45.3	43.2	57.1	43.7
SiamMOT	52.6	52.6	67.2	57.9
FairMOT	59.3	60.2	60.0	53.7

(b) Results for static-to-slow vs medium-to-fast moving tracks.

Method	IDF1(↑)		MOTA(↑)	
	long	short	long	short
CenterTrack	32.5	51.2	37.6	59.6
SiamMOT	39.8	59.6	51.4	70.3
FairMOT	44.3	68.3	45.2	64.4

(c) Results for tracks with medium-to-long vs short-to-no occlusions.

Method	IDF1(↑)			MOTA(↑)		
	s	m	l	s	m	l
CenterTrack	42.9	47.4	41.4	-3.9	51.9	62.2
SiamMOT	51.3	54.4	50.7	16.4	62.1	73.0
FairMOT	52.3	61.7	58.9	13.2	58.2	64.6

(d) Results for tracks with short (s), medium (m) and long (l) duration.

Table 3: Result comparison of models on different subsets of person tracks.

moving (e.g. waiting for the pedestrian green light). We find that the presence of such objects can inflate the evaluation metrics given the fact that tracking static objects is perceptually easier than tracking moving ones. This is because static objects do not require sophisticated motion models and do not exhibit any change in appearance over time unless they are occluded. In Tab. 3b, we show the performance for static vs. moving objects. Overall, MOTA is significantly higher for static tracks than for moving tracks, which indicates that static/slow-moving people are easier to be detected in our dataset. However, IDF1 performance is similar for both set of tracks, which suggests that the person's motion velocity is not strongly correlated with of its level of tracking difficulty level in our dataset.

Tracks with Full Occlusion. Being able to track such scenarios is of great importance in real-world tracking applications, especially when the camera is close to the ground where person-to-person occlusion is common. Tracking through full occlusion and keeping its identity unchanged is challenging in particular in video sequences where a large number of people are present. To this end, we report results on tracks with short-to-no occlusion and with medium-to-long occlusion, which are defined in Sec. 5.2. As shown in Tab. 3c, both the MOTA and IDF1 are substantially lower for tracks with medium-to-long occlusion. In this case, people are more likely to be partially occluded, which leads to more detection failures that contributes to lower MOTA. The huge gap in terms of IDF1 for all models suggests that preserving the same identity before and after the person is occluded is challenging and we hope that future research can improve performance for online trackers to track through long occlusion.

Track Duration. In Tab. 3d, we show the break-down results for tracks with short, medium and long duration, as defined in Sec. 5.2. There are a few interesting observations: 1) the IDF1 for long-duration tracks is the lowest, despite the fact that its corresponding MOTA is the highest. We find that this happens because long-duration tracks usually appear in high-angle view cameras (e.g. MEVA [17], Virat [40]) in our dataset, therefore detecting person in those videos is easier, which positively correlates with a higher MOTA; 2) the MOTA for short-duration track is abysmal, although it has a decent IDF1. We notice that the presence of short tracks are correlated with various challenging occlusion scenarios, for example, short tracks are associated to people in large crowds or people walking behind various objects (trees, vehicles), where the people are first visible, then become partially-occluded and disappear quickly. The challenges presented in short, medium, and long tracks are diverse and depending on the application each could be important. Thus we hope that researchers will adopt the practice of reporting metrics on these three categories separately in the future to give further insight into their model performance.

In summary, our dataset provides interesting and challenging cases for real-world tracking that includes various duration tracks, tracks with medium-to-long occlusion and small-size person tracks, on which existing state-of-the-art online trackers struggle.

7 Conclusion and Discussion

In this paper, we introduced a large scale real-world multi-person tracking dataset. The dataset is meticulously curated by (1) sourcing a set of videos that are diverse in terms of people density, camera angles, weather and scenery types as well as lighting conditions and (2) exhaustively annotating all persons in every frame with rigorous annotation and verification protocol that accommodates robust edge case handling. We demonstrated the value of the dataset by comparing it against existing datasets including MOT17 [39], HiEve [35], and MOT20 [20]. Our dataset is a magnitude larger than the most popular MOT17 dataset in terms of unique person tracks, number of videos, and total video duration. We further performed in-depth analyses of existing state-of-the-art online trackers on our dataset and observed interesting cases where current online trackers fail to perform well. We hope that the publication of this dataset will spark a new wave of research towards developing more usable tracking models in real-world multi-person tracking.

Socially responsible usage of the dataset. This dataset should primarily be used to improve person tracking algorithms, which can have a significant positive effect on many real-world video understanding problems including for example self-driving cars and human activity understanding. We ask the users of this dataset to use the data in a socially responsible manner, and request to not use the data to identify or generate biometric information of the people in the videos.

References

1. Fillerstock, https://fillerstock.com/ 5
2. Pexels, https://www.pexels.com/ 5
3. Pixabay, https://pixabay.com/ 5
4. Bai, H., Cheng, W., Chu, P., Liu, J., Zhang, K., Ling, H.: Gmot-40: A benchmark for generic multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6719–6728 (2021) 3
5. Beddiar, D.R., Nini, B., Sabokrou, M., Hadid, A.: Vision-based human activity recognition: a survey. Multimedia Tools and Applications **79**(41), 30509–30555 (2020) 3
6. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. EURASIP J. Image Video Process. **2008** (2008). https://doi.org/10.1155/2008/246309, https://doi.org/10.1155/2008/246309 11
7. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: European conference on computer vision. pp. 850–865. Springer (2016) 4
8. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016) 1, 4
9. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020) 3
10. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 4
11. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019) 1, 2
12. Chandrajit, M., Girisha, R., Vasudev, T.: Multiple objects tracking in surveillance video using color and hu moments. Signal & Image Processing: An International Journal (SIPIJ) **7**(3), 16–27 (2016) 3
13. Chandrakar, R., Raja, R., Miri, R., Sinha, U., Kushwaha, A.K.S., Raja, H.: Enhanced the moving object detection and object tracking for traffic surveillance using rbf-fdlnn and cbf algorithm. Expert Systems with Applications **191**, 116306 (2022) 3
14. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8748–8757 (2019) 3
15. Chang, S., Yuan, L., Nie, X., Huang, Z., Zhou, Y., Chen, Y., Feng, J., Yan, S.: Towards accurate human pose estimation in videos of crowded scenes. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4630–4634 (2020) 3
16. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) 3

17. Corona, K., Osterdahl, K., Collins, R., Hoogs, A.: Meva: A large-scale multiview, multimodal video dataset for activity detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1060–1068 (January 2021) 2, 5, 14
18. Datta, A., Shah, M., Lobo, N.D.V.: Person-on-person violence detection in video data. In: Object recognition supported by user interaction for service robots. vol. 1, pp. 433–438. IEEE (2002) 3
19. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: European conference on computer vision. pp. 436–454. Springer (2020) 3
20. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020) 1, 3, 7, 10, 11, 14
21. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009) 1, 2
22. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 304–311. IEEE (2009) 3
23. Ess, A., Schindler, K., Leibe, B., Van Gool, L.: Object detection and tracking for autonomous navigation in dynamic environments. The International Journal of Robotics Research **29**(14), 1707–1725 (2010) 3
24. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (Jun 2010) 2
25. Fabbri, M., Brasó, G., Maugeri, G., Ošep, A., Gasparini, R., Cetintas, O., Calderara, S., Leal-Taixé, L., Cucchiara, R.: Motsynth: How can synthetic data help pedestrian detection and tracking? In: International Conference on Computer Vision (ICCV) (2021) 3
26. Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., Cucchiara, R.: Learning to detect and track visible and occluded body joints in a virtual world. In: European Conference on Computer Vision (ECCV) (2018) 3
27. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 3
28. Guo, D., Wang, J., Cui, Y., Wang, Z., Chen, S.: Siamcar: Siamese fully convolutional classification and regression for visual tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6269–6277 (2020) 11
29. Han, X., You, Q., Wang, C., Zhang, Z., Chu, P., Hu, H., Wang, J., Liu, Z.: Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark (2021) 3
30. Held, D., Thrun, S., Savarese, S.: Learning to track at 100 fps with deep regression networks. In: European conference on computer vision. pp. 749–765. Springer (2016) 11
31. Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., Omari, S., Iglovikov, V., Ondruska, P.: One thousand and one hours: Self-driving motion prediction dataset. arXiv preprint arXiv:2006.14480 (2020) 3
32. Leal-Taixé, L., Canton-Ferrer, C., Schindler, K.: Learning by tracking: Siamese cnn for robust target association. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 33–40 (2016) 4

33. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4282–4291 (2019) 4
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 1, 2
35. Lin, W., Liu, H., Liu, S., Li, Y., Qian, R., Wang, T., Xu, N., Xiong, H., Qi, G.J., Sebe, N.: Human in events: A large-scale benchmark for human-centric video analysis in complex events. arXiv preprint arXiv:2005.04490 (2020) 3, 7, 10, 11, 14
36. Liu, W., Bao, Q., Sun, Y., Mei, T.: Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective. arXiv preprint arXiv:2104.11536 (2021) 3
37. Manen, S., Gygli, M., Dai, D., Gool, L.V.: Pathtrack: Fast trajectory annotation with path supervision. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 290–299 (2017) 2, 5
38. Mathur, G., Somwanshi, D., Bundele, M.M.: Intelligent video surveillance based on object tracking. In: 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE). pp. 1–6. IEEE (2018) 3
39. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016) 1, 3, 4, 7, 8, 9, 10, 11, 14
40. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR 2011. pp. 3153–3160. IEEE (2011) 5, 14
41. Pang, B., Li, Y., Zhang, Y., Li, M., Lu, C.: Tubetk: Adopting tubes to track multi-object in a one-step training model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6308–6318 (2020) 1
42. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 164–173 (2021) 3, 4
43. Rangesh, A., Trivedi, M.M.: No blind spots: Full-surround multi-object tracking for autonomous vehicles using cameras and lidars. IEEE Transactions on Intelligent Vehicles **4**(4), 588–599 (2019) 3
44. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015) 4, 11, 12
45. Rezaei, M., Azarmi, M., Mir, F.M.P.: Traffic-net: 3d traffic monitoring using a single camera. arXiv preprint arXiv:2109.09165 (2021) 3
46. Ristani, E., Solera, F., Zou, R.S., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II. Lecture Notes in Computer Science, vol. 9914, pp. 17–35 (2016). https://doi.org/10.1007/978-3-319-48881-3_2, `https://doi.org/10.1007/978-3-319-48881-3_2` 11
47. Ristani, E., Tomasi, C.: Features for multi-target multi-camera tracking and re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6036–6046 (2018) 4
48. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018) 11

49. Shuai, B., , Li, X., Kundu, K., Tighe, J.: Id-free person similarity learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022) 4
50. Shuai, B., Berneshawi, A., Li, X., Modolo, D., Tighe, J.: Siammot: Siamese multi-object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12372–12382 (2021) 1, 3, 4, 11, 12
51. Song, L., Yu, G., Yuan, J., Liu, Z.: Human pose estimation and its application to action recognition: A survey. Journal of Visual Communication and Image Representation **76**, 103055 (2021) 3
52. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020) 3
53. Sundararaman, R., De Almeida Braga, C., Marchand, E., Pettre, J.: Tracking pedestrian heads in dense crowd. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3865–3875 (2021) 3
54. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019) 4
55. Wang, G., Wang, Y., Zhang, H., Gu, R., Hwang, J.N.: Exploit the connectivity: Multi-object tracking with trackletnet. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 482–490 (2019) 4
56. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: European Conference on Computer Vision. pp. 107–122. Springer (2020) 4
57. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017) 1, 4
58. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.M.: A scalable approach to activity recognition based on object use. In: 2007 IEEE 11th international conference on computer vision. pp. 1–8. IEEE (2007) 3
59. Xu, J., Cao, Y., Zhang, Z., Hu, H.: Spatial-temporal relation networks for multi-object tracking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3988–3998 (2019) 4
60. Xu, Y., Osep, A., Ban, Y., Horaud, R., Leal-Taixé, L., Alameda-Pineda, X.: How to train your deep multi-object tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6787–6796 (2020) 4
61. Yiyang Gan, Ruize Han, L.Y.W.F.S.W.: Self-supervised multi-view multi-human association and tracking. In: ACM MM (2021) 3
62. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020) 3
63. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015) 1
64. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2403–2412 (2018) 11

65. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021) 1, 4, 11, 12
66. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision. pp. 474–490. Springer (2020) 1, 4, 10, 11, 12
67. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. In: arXiv preprint arXiv:1904.07850 (2019) 4, 10, 11, 12