SoccerNet-Tracking: Multiple Object Tracking Dataset and Benchmark in Soccer Videos

Anthony Cioppa[1]* Silvio Giancola[2]* Adrien Deliège[1]* Le Kang[3]* Xin Zhou[3]*
Zhiyu Cheng[3] Bernard Ghanem[2] Marc Van Droogenbroeck[1]
[1] University of Liège [2] KAUST [3] Baidu Research

Abstract

Tracking objects in soccer videos is extremely important to gather both player and team statistics, whether it is to estimate the total distance run, the ball possession or the team formation. Video processing can help automating the extraction of those information, without the need of any invasive sensor, hence applicable to any team on any stadium. Yet, the availability of datasets to train learnable models and benchmarks to evaluate methods on a common testbed is very limited. In this work, we propose a novel dataset for multiple object tracking composed of 200 sequences of 30s each, representative of challenging soccer scenarios, and a complete 45-minutes half-time for long-term tracking. The dataset is fully annotated with bounding boxes and tracklet IDs, enabling the training of MOT baselines in the soccer domain and a full benchmarking of those methods on our segregated challenge sets. Our analysis shows that multiple player, referee and ball tracking in soccer videos is far from being solved, with several improvement required in case of fast motion or in scenarios of severe occlusion.

1. Introduction

Imagine you are scouting a new striker for your soccer team. How would you evaluate the skills of all potential candidates? Prior information on the scouted players are of paramount importance. In practice, several metrics typically supports the scouting choices. For instance, player endurance along a full game, total distance run, top speed in counter-attacks, number of ball possessions, assists or goals are only few examples of characteristics that your team would consider before hiring your next soccer talent. Yet, all those important statistics require the information of the players position in time, also known as player tracking.

The increasing demand from sports professionals for more and more advanced analytics calls for automatic video processing. A low-level understanding of the game and their actors in soccer videos could provide such players and team statistics. In particular, Multiple Object Tracking (MOT) methods applied to soccer videos could identify the position and trajectory of each player in a video, as well as the ball or any other important actor. To achieve this objective, visual computing appears handy since it does not require any invasive sensor placed on the players. Indeed, visual tracking extracts valuable information for each player, such as its positioning in particular scenarios, speed with and without the ball on his feet, running analysis, *etc.* as those information can indicate a positive outcome such

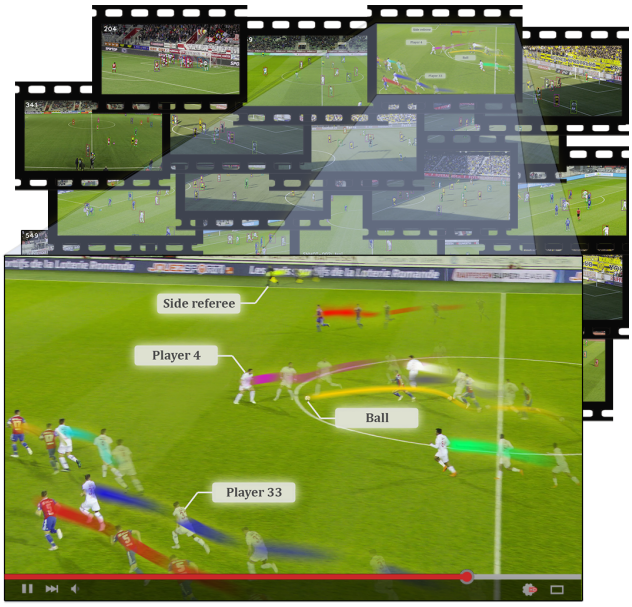

Figure 1. **SoccerNet-Tracking.** We propose a novel dataset for Multiple Object Tracking (MOT) in soccer videos including the players, the ball, and the referees. Our dataset is composed of 200 sequences of 30s each, representative of interesting moments from 12 soccer games, densely annotated with player tracklets, teams and jersey numbers. Moreover, we also include a fully annotated 45min half time video, focusing on long-term tracking.

(*) Equal contributions. Data/code available at www.soccer-net.org.

as a goal, for instance. Tracking also gathers valuable information for the team as a whole as it helps understanding specific game strategies, and assists in proposing a counter strategy against an adversary team.

However, several challenges arise in soccer player tracking: **(i)** players appear visually very similar between each others in soccer videos, with only a few characteristics to differentiate players from the same team (typically jersey number, or shoes color). **(ii)** Players are often occluding each others in specific game scenarios (*e.g*. corners), increasing the difficulty of recognizing them and with a risk of switching their identities. **(iii)** Tracking the soccer ball is extremely challenging due to its small size (*e.g*. < 100 pixels), occlusions from players, fast motion, incurring blurring effects, and shape shifting on the video frames. Yet, several downstream tasks can benefit from soccer player tracking techniques. For *action spotting*, player trajectories in the field are highly indicative of the nature of the event occurring (*e.g*. crowd around goal when corners happen, players running after a goal occurs, *etc*.). For *highlight generation*, an incredible performance from a player can be retrieved by analysing the trajectories of the player and the ball (*e.g*. an impressive sprint or dribbling of several players, a ball crossing the goal line or going out the field, *etc*.).

In this work, we propose SoccerNet-Tracking, a large-scale dataset and benchmark for Multiple Object Tracking (MOT) in soccer videos, with a focus on players, balls and referees, as illustrated in Figure 1. We provide 200 tracking sequences gathered around 11 interesting classes of soccer actions (*e.g*. goal, corner, direct free-kick, foul, *etc*.) and corresponding to challenging tracking scenario. Furthermore, we have annotated a complete 45-minute half-time of densely annotated soccer video to evaluate long-term tracking. We dedicate part of the sequences for training and benchmarking purposes; annotations for the remaining sequences are kept segregated for future open challenges to prevent any over-fitting. Finally, we benchmark state-of-the-art MOT baselines on our novel dataset, and run an extensive analysis of challenging tracking scenarios.

Contributions. We summarize our contributions as follows. **(i)** We propose the largest dataset for multi-object tracking in soccer videos. It is composed of 200 sequences of 30s each, fully annotated with bounding boxes and ID tracklets at 25fps, and a complete 45-minutes half-time for long-term tracking. **(ii)** We propose an extensive benchmark of the most recent multi-object trackers on our new dataset, highlighting the difficulties of soccer players tracking in different scenarios, and providing a first state of the art for the tracking task.

2. Related Work

Our work relates to datasets and methods for multiple object tracking and its application to sports analysis.

2.1. Multiple Object Tracking (MOT)

Tracking objects in videos consists in localizing and following objects of interest across video sequences [49].

Methods. The task of Multiple Object Tracking (MOT) is often approached with a tracking-by-detection paradigm [5, 9, 67, 80, 81]. It consists in localizing objects of interest at each frame, and associating them into tracklets by finding consecutive correspondences in time. Object detection boasts a large active research community; as a result, the research interests in MOT lie in identifying object tracklets out of those detections, handling object disappearances with re-identification techniques and detector failures with temporal interpolation. Bewley *et al*. [8] proposed **SORT** that leverages Kalman filtering with an Hungarian algorithm to associate overlapping bounding boxes. The extension **DeepSORT** [17] incorporates deep appearance features into the association metric. Bergmann *et al*. [5] proposed **Tracktor**, a method that exploits the regression head of object detection models to perform temporal realignment. Zhang *et al*. [81] proposed **FairMOT**, that fine-tunes the detection model aside with the re-identification module that associates new detected objects with the previous list of tracklets. Zhang *et al*. [80] proposed **ByteTrack**, that considers every detection boxes despite their confidence score, relying on Kalman filtering for the association task. In this work, we apply the latest research in MOT to soccer videos.

Datasets. The progress in MOT would not have been possible without the numerous datasets released publicly to the community. The first attempts in building MOT datasets surged from PETS2009 [26] and TUD [1], yet they are relatively small in scale. MOT [19, 41, 52], KITTI-T [30] and DETRAC [50, 74] made tremendous efforts in providing curated video sequences annotated with bounding boxes associated in tracklets for the training and evaluation of MOT methods. Those datasets contain different levels of difficulty, from challenging illumination to several occlusions. Yet, all those datasets focus on tracking pedestrians and/or vehicles for autonomous navigation, traffic monitoring and security purposes. Only a few recent works targeted different domains ranging from human faces [68], animals [55, 59], biological cells [2] up to more generic object [16]. In this work, we provide a large scale dataset for MOT in a novel domain, in particular soccer videos.

2.2. Sports Video Understanding

Sports videos are investigated for different semantic understandings, ranging from high-level temporal action localization down to low-level player detection and tracking. Recent years have witnessed a surge in video understanding, with datasets and tasks ranging from video classification [28, 39, 62], action detection [24, 32, 36, 44, 64, 70], camera shot segmentation [18, 37], and highlight summa-

rization [12]. The SoccerNet series [11, 18, 32] provides the largest datasets for soccer broadcast understanding, with comprehensive benchmarks for Action Spotting, Replay Grounding, Player Re-Identification and Pitch Localization. In this work, we complement the SoccerNet effort with the task of multiple object tracking. We propose a novel dataset of 200 tracking sequences and a complete 45-minutes half-time for long-term tracking, fully annotated with tracklets of players, referees and balls, meeting the SoccerNet standards in terms of data distribution and benchmark.

2.3. Player Localization and Tracking in Sports.

Sports videos are also investigated at a player level, *e.g.* retrieving player's jersey [31, 42, 43] or team [34, 40], localizing their position in the field [57, 60, 77], estimating their motion [51] or forecasting their intention [3, 71].

Player Localization. In this literature, Rao *et al.* [57] presented a pre-deep learning algorithm that detects players using color gradients and ground lines detected with Hough transform. Follow up works adapted the latest advances in generic object detection [58] for soccer broadcast understanding. Nekoui *et al.* [53] investigated sports athletes under challenging positions and Liu *et al.* [45] improved object localization by learning their relationships. Istasse *et al.* [34] and Koshkina *et al.* [40] learned to discriminate player teams in unsupervised fashions. Cioppa *et al.* [14] distill, in an online fashion, a segmentation network trained on generic data for single camera soccer videos, and a detection network for real-time player detection in amateur sports [15]. Sanford *et al.* [61] and Cioppa *et al.* [13] both leverage player localization for action recognition and spotting. Inhere, we tackle localization in time, *i.e.* tracking.

Player Tracking. Tracking players in time is extremely useful to gather per-player statistics. The literature in MOT is extensive. Manafifard *et al.* [51] provide a comprehensive survey on player tracking in soccer videos. Iwase *et al.* [35] proposed an elegant solution to track players using a background subtraction method, a triangulation technique from multiple cameras to project their positions on the pitch, and a Kalman filter to identify their trajectories. Figueroa *et al.* [27] and Sullivan *et al.* [66] proposed similar pipelines by considered player positions as nodes in graphs, and trajectories as edges between nodes. Nillius *et al.* [54] presented a Bayesian framework to link identities, and Xing *et al.* [76] included a progressive observation modeling process. Lu *et al.* [46] proposed a learning approach to identify and track players in videos, based on handcrafted visual features and Kalman filters. More recently, Hurault *et al.* [33] transferred an object detection model trained on generic objects, fine-tuned for soccer player detection and tracking in a self-supervised fashion. Yet, those methods leverage small-scale and private datasets. With this work, we release the largest ever dataset for MOT in sports, enabling a fair benchmark for future development in this community.

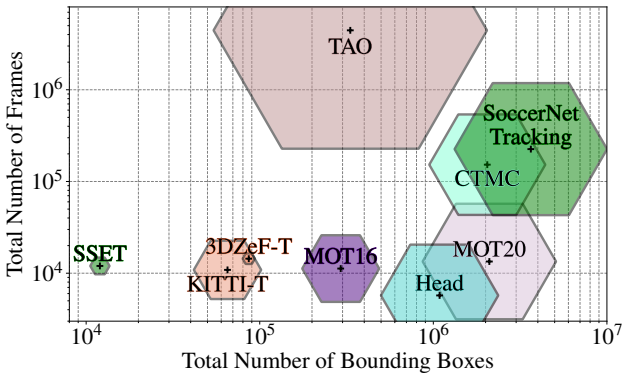

Figure 2. **SoccerNet-Tracking against other tracking datasets.** The area is proportional to the number of unique tracklets in each dataset. SoccerNet-Tracking offers a great trade-off between the total number of bounding boxes, frames and unique tracklets. Furthermore, only SSET proposes soccer sequences, which is much smaller and only focuses on single object tracking.

Dataset. Most works in soccer player MOT benchmark their method on small-scale proprietary datasets [7, 72]. Pettersen *et al.* [56] shared videos and GPS tracker logs acquired at the **Aflheim** Stadium in Norway. Yet, they provide the GPS logs and videos for only 3 games. The closest effort to our work originated from Yu *et al.* [78] and includes the **SSET** dataset [25] and the **BSPT** baseline [65]. SSET is composed of 282 hours of video, from which 80 tracking sequences are extracted. Because SSET relies on broadcast videos, they first identify interesting far-away camera shots. As a results their sequences only lasts about $10s$ (248 frames at 25fps). Moreover, SSET focuses on Single Object Tracking rather than Multiple Object Tracking, where a single player of interest is tracked, initialized with its bounding box on the first frame. Differently, our videos originate from single cameras following the action, and our 200 sequences are longer (30 seconds) and handpicked to represent specific scenarios of interest in a soccer game, representative of interesting game events and tracking scenarios. Furthermore, our densely annotated 45-minutes half-time sequence is the first public release of long-term tracking data in the sports community.

3. SoccerNet-Tracking Dataset

Data collection. Our SoccerNet-Tracking sequences consist of main-camera videos from 12 complete soccer games recorded during the 2019 Swiss Super League. All videos are captured at $1080p$ resolution (Full-HD) and provided at 25 frames per second. These single-camera sequences are particularly suited for a tracking task compared with broadcast videos typically found in soccer datasets, where cam-

Dataset	Sequences	Frames	Tracklets	Bounding boxes	Domain	Task
MOT16 [52]	14	11,235	1,276	292,733	Pedestrians	MOT
MOT20 [19]	8	13,410	3,833	2,102,385	Urban (crowded)	MOT
KITTI-T [30]	50	10,870	977	65,213	Autonomous Driving	MOT
Head [68]	5	5,723	2,965	1,086,790	Pedestrian (heads)	MOT
TAO [16]	3	4,447,038	16,104	332,401	Generic	MOT
3DZeF-T [55]	8	14,398	32	86,452	Fish	3D MOT
CTMC [2]	86	152,498	2,900	2,045,834	Cells	MOT
SSET [25]	80	12,000	80	12,000	Soccer	SOT
SN-Tracking (ours)	201	225,375	5,009	3,645,661	Soccer	MOT

Table 1. **Comparison of SoccerNet-Tracking with other tracking datasets.** Our dataset contains the largest set of bounding boxes and sequences across all tracking dataset, as well as the second most number of frames and unique tracklets. This shows that SoccerNet-tracking is a great dataset for research in tracking. Also, our dataset is the first multi-object tracking dataset in soccer, scaling by a large factor the previous SSET [25] dataset that only focused on single-object tracking.

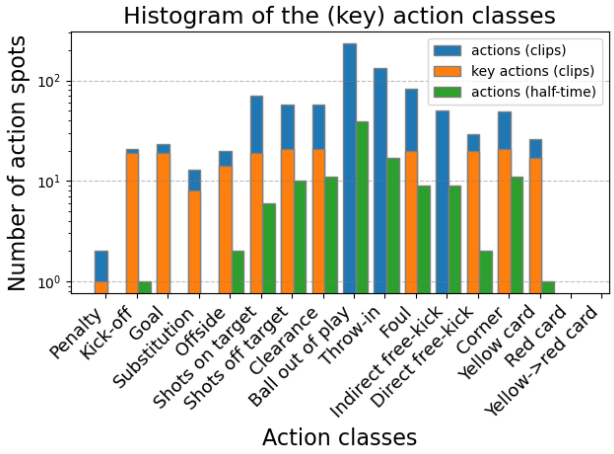

Figure 3. **Distribution of the action classes.** Number of action classes within the 200 clips and the whole half-time separately. For the 200 clips, the key action distribution correspond to the anchoring actions in the clip selection process. Note that within the 12 games, we have no red card or yellow to red card events.

era changes and replays break the continuity of the tracks. Since manually annotating the entire 12 games for tracking would be too costly, we select a subset of these games by extracting 30-seconds clips at interesting moments in the games, with an additional complete half-time to be annotated. The first step is therefore to find these interesting moments in the videos. For that, we start by annotating all events occurring in the game following the same process and action classes than the action spotting task of SoccerNet-v2 [18]. This resulted in 3,132 extracted action spots corresponding to 17 action classes among 1,210 minutes of video. We noticed that some classes such as for corners or cards are particularly challenging for a tracking task as they often involve clustered players with a lot of occlusions. Based on these annotations, we select the entire half-time to annotate as the one containing the most challenging action spots, while keeping good action diversity. Then, among the remaining 11 games, we select 200 30-seconds clips around key action spots, corresponding to challenging events for tracking, uniformly sampled across 11 action classes. Of course, more than one action may happen within these 30-seconds clips, bringing diversity in player configurations and movements. Finally, we ensure that two clips never overlap to avoid redundancy. All-in-all, the selected data amounts to 150,000 frames for the clips and 75,375 frames for the whole half-time, totalling 225,375 frames annotated with tracking information, as shown in Table 1 and illustrated in Figure 2. The action class distribution among the video clips and the half-time is given in Figure 3.

Tracking annotations. The tracking information is annotated on SuperAnnotate [69], a professional platform specialized in data annotation. We first define 5 classes of objects of interest to track, corresponding to the main actors of a soccer game: player, goalkeeper, referee, ball, and other (*e.g.* medical staff or coaches entering the field). For the players and goalkeepers, we annotate an extra team tag specifying the side of the team (left or right) as well as the jersey number when visible at least once in the video. We also further refine the referee annotation between main referee, side top referee and side bottom referee. Even though these extra metadata are not used for the tracking task described in this paper, they might be useful for further work on team assignment and jersey number recognition. All bounding boxes are manually annotated as tight as possible around the object of interest at specific key frames. In between key frames, the bounding boxes are interpolated both in their position and size using linear interpolation. On average, only 11.5% of the bounding box are interpolated in the videos, corresponding to very dense manual annotations. Finally, the bounding boxes are assigned a unique

Figure 4. **Example of tracking annotations in our dataset for challenging events.** From top to bottom: (a) Corner actions often display a lot of occluded and clustered players. (b) Direct free-kicks also show clustered players going in the same direction with many crossings between players. (c) Penalties display almost all objects moving in the same direction often followed by cheering. (d) Shot on target actions involve high speed movement of the ball and players towards the goal.

Object Class	Unique tracklets	Bounding boxes
Player	4,005	2,992,173
Goalkeeper	262	130,109
Referee	432	301,025
Ball	297	215,156
Other	13	7,198
Total	5,009	3,645,661

Table 2. **Object statistics.** Players are way more represented both in terms of unique tracklets and bounding boxes than other classes.

track identifier (ID). In total, we have annotated around $3.6M$ bounding boxes corresponding to $5{,}009$ unique objects, with 96% of players and goalkeepers having a jersey number assigned. Table 2 provides further statistics about the object class distribution.

Novelty. Unlike traditional tracking datasets, we purposely choose to keep the same ID for an object that leaves the camera frame and comes back at a later time during the same clip. This makes our setup much closer to real-world soccer application, where identifying and tracking a player are key components for analyzing his performances. Most current trackers do not propose such long-term re-identification, making our dataset a perfect sandbox for pushing research towards long-term object re-identification in tracking. Furthermore, the players of a same team have very similar appearances, making the task particularly challenging in case of crossings between players. Some examples of tracking annotations may be found in Figure 4 for several hard cases, including player clusters and crossing between players. SoccerNet-Tracking is among the largest tracking datasets publicly available, and the largest tracking dataset related to sports. Table 1 and Figure 2 compare SoccerNet-Tracking with the other tracking datasets. As can be seen, our dataset contains the most bounding boxes and sequences, as well as the second most number of frames and unique tracklets. Furthermore, it is the first MOT dataset in soccer, supplanting the previous SSET [25] dataset that only focused on single-object tracking.

Data format. The 12 games are split equally into 4 sets: train, test, and two challenge sets which are kept private at the moment. In particular, this accounts for 57 30-seconds clips for the train set, 49 clips for the test set, 58 clips for our first public challenge, and 37 clips for our second challenge, including the entire half-time video in the latter. Then, the

folder and data structure are chosen to be as close as possible to the MOT20 [19] format. We believe that uniformity between datasets is valuable for the tracking community, especially to benchmark new methods. In particular, one can use the evaluation and visualization kits of MOT such as *TrackEval* [47] and *MOTChallengeEvalKit* [20] on our data. For the sake of completeness, we detail the data format in the following, highlighting the slight non-disruptive discrepancies with the MOT20 dataset. Each set is separated in its own folder containing all of its sequences, also split in separate sub-folders named after the sequence name. The list of sequences in a particular set may be found in the *seqmaps* folder, following the MOT20 convention. For each sequence, images are extracted from the video and converted to JPEG files named using the frame ID (for instance from *000001.jpg* to *000750.jpg* for the 30-seconds clips). The ground truth and detections are stored in their own subfolder in comma-separate csv files with 10 columns. These values correspond in order to: frame ID, track ID, top left coordinate of the bounding box, top y coordinate, width, height, confidence score for the detection (always 1. for the ground truth) and the remaining values are set to -1 as they are not used in our dataset, but are needed to comply with the MOT20 requirements. We also provide two configuration files for each sequence. The first one, *seqinfo.ini*, provides information about the video format, such as the length and resolution. The second configuration file, *gameinfo.ini*, provides information about the events and objects in the video sequence such as the anchoring main event and the position of the clip within the whole game. It also provides metadata about the tracks such as the precise object class and jersey number when available (otherwise, the tag is a generic letter used to uniquely identify the object). Note that unlike MOT20, we provide the ground-truth annotations for the test set publicly, so that researchers can benchmark their results locally without relying on an external evaluation server. However, the ground-truth data for both challenges are kept private and the evaluation can only be done on our submission platform with limited daily and monthly submissions to prevent overfitting.

4. Benchmarks

Task. Multi-object tracking (MOT) aims at recovering trajectories of multiple objects in time by estimating object bounding boxes and identities in videos sequences. In our case, the objects of interest include players, goalkeepers, referees, balls, and other actors such as the medical staff or the coaches entering the field. In this work, we consider two tasks: (1) a pure re-identification task that considers ground-truth detections, or (2) a complete tracking task that expects detecting the objects of interest from the raw video.

Baselines. We evaluate three state-of-the-art MOT methods. **(i)** *DeepSORT* [17] is an extension of SORT [8] that performs Kalman filtering [38] on detected objects and applies the Hungarian algorithm with an association metric that measures bounding box overlap. DeepSORT extends SORT by incorporating appearance features generated by a deep convolutional neural network into the association metric. **(ii)** *FairMOT* [81] performs both object detection and re-identification feature generation in a single shot, similar to JDE [73], but aims for a good balance between the two tasks. **(iii)** *ByteTrack* [80] proposes an effective and generic association method that tracks objects by associating every detection box instead of only the high score ones. For the low score detection boxes, ByteTrack uses IOU scores as their similarities to assign tracklets and recover true objects or filter out the background detections. For the detection part, ByteTrack leverages the new YOLOX [29] detector. It is also the current state of the art on the MOT20 dataset.

Metrics. To evaluate the different aspects of tracking, we consider two main metrics: MOTA from the CLEAR metrics [6], and the more recent metric HOTA [48]. The MOTA [6] metric has been widely used as the main metric for many MOT benchmarks. However, it focuses more on the detection performance and weighs significantly less on the association performance. To circumvent this limitation, Liuten *et al*. [48] disentangle the performances for detection (DetA) and association (AssA) and combine both in a single HOTA metric. More details on the HOTA metric can be found in their paper [48]. For our benchmark, we consider HOTA as the main metric. We evaluate the above baseline methods on our SoccerNet-Tracking test set, and show the results in Table 3. In the "Setup" column, "w/ GT" indicates that ground-truth detections are provided to the baselines, while "w/o GT" indicates the more challenging setting, *i.e*. each algorithm uses its own detector. Therefore, in the "w/ GT" setup we are able to focus our study only on the association performance, given the same perfect detection results for every algorithm.

Implementation details. For both DeepSORT and FairMOT, we use the implementations from PaddleDetection [4], with input dimension 1088×608. Specifically, the DeepSORT model uses "JDE YOLOv3" for object detection and "PCB pyramid" for extracting re-ID features (see [17] for more details). The JDE YOLOv3 detector is pretrained for 30 epochs on the "MIX" dataset, which is a collection of six datasets including Caltech Pedestrian [21], CityPersons [79], CUHK-SYSU [75], PRW [82], ETHZ [22], and MOT17 [52]. The FairMOT model uses the DLA-34 backbone, also pretrained on the same "MIX" dataset for 30 epochs, as described in [23]. We also finetune the pre-trained FairMOT model on our train set for 10 epochs, and denote the resultant model FairMOT-ft. For ByteTrack, we use the open source code [10] provided by the authors. We use the pre-trained model "byte-

Algorithm	Setup	HOTA	DetA	AssA	MOTA
DeepSORT	w/ GT	69.552	82.628	58.668	**94.844**
FairMOT	w/ GT	-	-	-	-
ByteTrack	w/ GT	**71.500**	**84.342**	**60.718**	94.572
DeepSORT	w/o GT	36.663	40.022	33.759	33.913
FairMOT	w/o GT	43.911	46.317	41.778	50.698
ByteTrack	w/o GT	47.225	44.489	50.257	31.741
FairMOT-ft	w/o GT	**57.882**	**66.565**	**50.492**	**83.565**

Table 3. **Leaderboard**. Evaluation of the state-of-the-art tracking methods on our new SoccerNet-Tracking test set with (w/) and without (w/o) ground-truth detections. DeepSORT, FairMOT and ByteTrack increasingly improve the performance, and fine-tuning FairMOT leads to the first state of the art on our new dataset.

track_x_mot20" trained on CrowdHuman [63] and MOT20 [19] datasets. The option for mixed precision evaluation (flag fp16) is turned on and "match_thresh" is set to 0.8. To use the pretrained model without ground-truth detections, the images are resized to 1600×896. With the ground-truth detections, we keep the original image size of 1920×1080. All other parameters are set to their default values.

Main Results. The results for the three baselines are presented in Table 3 for our two tracking setups, *i.e*. with and without ground-truth detections. First, *when ground-truth detections are provided*, ByteTrack slightly outperforms DeepSORT in AssA. Yet, they have similar MOTA scores, indicating that both algorithms perform quite similarly on the association task only. It is important to note that both algorithms do not achieve 100% in DetA, as common intuition would suggest. This is due to the fact that some of the detections get filtered out in the association process, especially in the case of fast motion. Let us note that we could not test FairMOT in this setup as its pipeline does not allow injecting external detections for the association task.

Second, *when ground-truth detections are not provided*, we can observe increasingly better performance from DeepSORT to FairMOT to ByteTrack, which is aligned with other MOT benchmarks. Furthermore, we note that FairMOT has better detection capabilities (DetA) than other baselines while ByteTrack has better association capabilities (AssA). Consequently, FairMOT has a better MOTA as this metric favors detection over association performance.

Third, we can see that fine-tuning FairMOT on our data improves both the detection and association performance, leading to state-of-the-art scores for both HOTA and MOTA. This improvement mainly comes from two sources: (i) The original detectors for the three baselines are trained to detect all humans in the video, hence including the audience, whereas our annotations only select players or staff on the field. Therefore, this results in lots of false positive detections outside the field that significantly lower the detection score. (ii) The original detectors are only trained to detect humans and hence never detect the ball which is considered in our annotations. Fine-tuning our model to only consider the humans and the ball on the field leads to less false positives and false negatives in the detections and therefore improves both the HOTA and MOTA scores.

Algorithm / Setup	ByteTrack - HOTA (DetA / AssA) w/ GT	ByteTrack - HOTA (DetA / AssA) w/o GT
Clearance	71.0 (82.2 / 61.7)	49.6 (48.9 / 54.7)
Corner	63.9 (83.8 / 49.6)	41.6 (45.5 / 39.7)
Direct free-kick	66.5 (84.5 / 53.2)	46.2 (55.1 / 39.5)
Foul	65.7 (82.6 / 53.3)	50.4 (54.8 / 46.9)
Goal	70.9 (82.2 / 61.2)	40.3 (29.6 / 54.8)
Kick-off	69.3 (84.5 / 57.1)	47.6 (49.8 / 46.3)
Offside	71.2 (83.3 / 61.1)	46.3 (42.7 / 53.4)
Penalty	70.4 (84.1 / 58.9)	31.0 (20.1 / 47.8)
Shots off target	74.0 (84.8 / 65.1)	49.2 (47.5 / 51.9)
Shots on target	67.6 (83.3 / 55.4)	46.6 (51.6 / 43.3)
Substitution	**84.2 (90.3 / 79.3)**	**64.7 (63.1 / 66.3)**
Yellow card	**74.1 (83.6 / 65.7)**	**53.5 (54.9 / 52.8)**

Table 4. **Per-class analysis.** Comparison of HOTA (DetA/AssA) for different event classes for ByteTrack w/ and w/o ground truth detections. The hardest events include goals, penalties, and corners with many clustered players, while the **easiest events** correspond to substitutions and yellow card with mostly static players.

Per-class analysis. Furthermore, we provide a thorough performance study per action class of the ByteTrack baseline in Table 4 for our two setups. *When ground-truth detections are provided*, detection scores are significantly lower for actions involving fast motion, mostly due to the filtering from the association process as explained above. For instance, substitutions are mostly static events and therefore have a high DetA, while clearances and goals involve fast moving players leading to lower DetA scores. This is also reflected in the overall HOTA score, for which the easiest classes are substitutions and yellow cards. Interestingly, the hardest cases are corners, which may be explained by lots of ID switches, as reflected by AssA, with players crossing each other multiple times during those scenarios.

Last, *when ground-truth detections are not provided*, substitutions and yellow cards remains the easiest scenarios to resolve, meaning that the generic detector already does a pretty good job in tracking the players in those scenarios. Corners and fouls remain challenging in this setup, but the two hardest scenarios are now goals and penalties, which often involve challenging dense player clusters such as celebrations or walls of players. In those cases, object detectors perform poorly due to players overlaps and occlusions. Indeed, we can see that when detections are given, these classes are actually well tracked, indicating that the difficulty is rather in detecting the objects in these scenarios than associating the correct bounding boxes together.

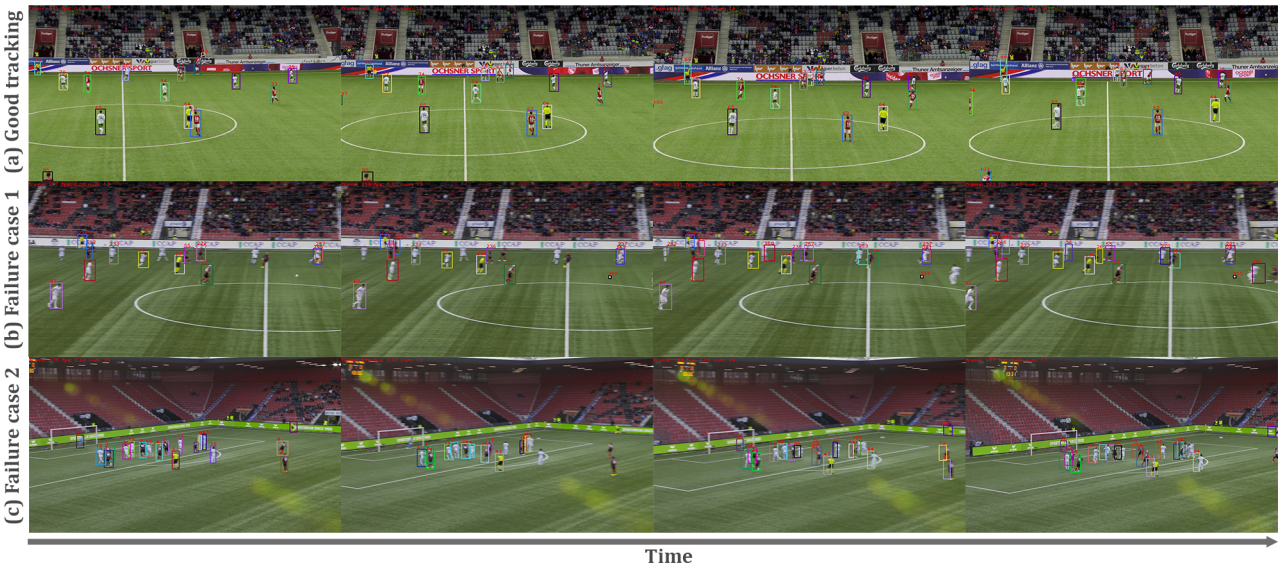

Figure 5. **Qualitative tracking results.** Tracking sequences produced by ByteTrack with ground-truth detections. Sequence (a) represents a good tracking of the players, even after some players or the referee are partially occluded. Sequence (b) shows an example of challenging association due to fast motion of the ball and players between consecutive frames. Sequence (c) displays a challenging free-kick scenario where many players are clustered together resulting in extreme occlusions and poor tracking results.

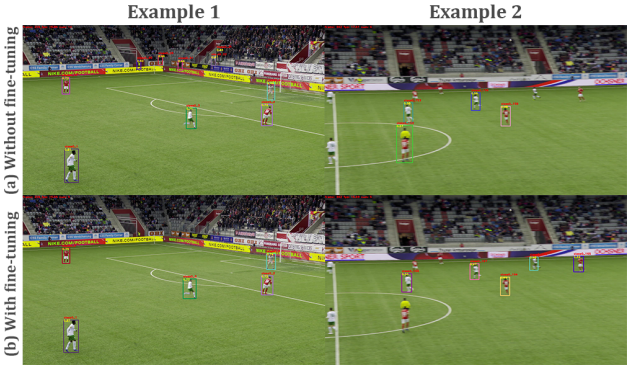

Figure 6. **Qualitative fine-tuned results.** Comparison of FairMOT predictions before and after fine-tuning on our dataset in the setup "w/o GT". Example 1 shows that the fine-tuned model does not consider the audience anymore in its predictions. Example 2 shows that we can also better detect player bounding boxes, despite the fact that fast motion blur remains challenging.

Qualitative results. We show three different sequences for a qualitative analysis of the ByteTrack in Figure 5. In sequence (a), we can see that the players are well tracked even after some players or the referee partially occlude each other. Sequence (b) shows an example of challenging association due to fast motion of the ball and players between consecutive frames. Sequence (c) shows a free-kick scenario, where many players are clustered, resulting in extreme occlusions and ByteTrack performing very badly. Our hope is that future methods have better re-identification capabilities to resolve hard cases like these.

Finally we analyse the effect of fine-tuning FairMOT on two frames in Figure 6. In the first example, we can see the inability of the pre-trained detector to distinguish between people in the audience and players on the field. In contrast, the fine-tuned detector has learned the semantic roles of each actor, discarding the audience. The second example shows a harder case with fast moving objects, where the fine-tuned detector learns to detect some of the fast moving players but not all of them. This indicates that further work is needed to improve detection performance as well.

5. Conclusion

We release the novel SoccerNet-Tracking dataset, which is the largest dataset for multi-object tracking in soccer, featuring challenges such as high-speed movements and highly occluded objects. This topic is important for many research and industrial application, which can have a direct impact on soccer. As a first approach, we study three state-of-the-art methods for multi-object tracking and discuss the challenging situations in our dataset that are not correctly tackled by those methods in their current state. With SoccerNet-Tracking, we have set a first benchmark and aim at pushing the computer vision community towards better tracking methods, including long-term re-identification in challenging environments, by organizing tracking challenges.

Acknowledgement. This work was supported by the Service Public de Wallonie (SPW) Recherche, under Grant No. 2010235 – ARIAC by DigitalWallonia4.ai, and KAUST Office of Sponsored Research (CRG2017-3405).

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3D pose estimation and tracking by detection. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 623–630, San Francisco, CA, USA, June 2010. 2

[2] Samreen Anjum and Danna Gurari. CTMC: Cell tracking with mitosis detection dataset challenge. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 982–983, Seattle, WA, USA, June 2020. 2, 4

[3] Adrià Arbués Sangüesa, Adrián Martín, Javier Fernández, Coloma Ballester, and Gloria Haro. Using player's body-orientation to model pass feasibility in soccer. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW)*, pages 3875–3884, Seattle, WA, USA, June 2020. 3

[4] PaddlePaddle Authors. PaddleDetection, object detection and instance segmentation toolkit based on PaddlePaddle. https://github.com/PaddlePaddle/PaddleDetection, 2019. 6

[5] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 941–951, Seoul, South Korea, Oct.-Nov. 2019. 2

[6] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP J. Image and Video Process.*, 2008:1–10, May 2008. 6

[7] Marco Bertini, Alberto Del Bimbo, and Walter Nunziati. Player identification in soccer videos. In *ACM SIGMM Int. workshop on Multimedia inf. retrieval*, pages 25–32, Nov. 2005. 3

[8] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 3464–3468, Phoenix, AZ, USA, Sept. 2016. Inst. Elect. and Electron. Engineers (IEEE). 2, 6

[9] Erik Bochinski, Tobias Senst, and Thomas Sikora. Extending IOU based multi-object tracking by visual information. In *IEEE Int. Conf. Adv. Video and Signal Based Surveillance (AVSS)*, pages 1–6, Auckland, New Zealand, Nov. 2018. 2

[10] ByteTrack Code. https://github.com/ifzhang/ByteTrack. 6

[11] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-v3: Scaling up SoccerNet with multi-view spatial localization and re-identification. *Submitted to Scientific Data*, 2022. 3

[12] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B Moeslund. A context-aware loss function for action spotting in soccer videos. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 13123–13133, Seattle, WA, USA, June 2020. 3

[13] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Floriane Magera, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. Camera calibration and player localization in SoccerNet-v2 and investigation of their representations for action spotting. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW), CVsports*, pages 4537–4546, Nashville, TN, USA, June 2021. 3

[14] Anthony Cioppa, Adrien Deliège, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. ARTHuS: Adaptive real-time human segmentation in sports through online distillation. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW), CVsports*, pages 2505–2514, Long Beach, CA, USA, June 2019. 3

[15] Anthony Cioppa, Adrien Deliège, Noor Ul Huda, Rikke Gade, Marc Van Droogenbroeck, and Thomas B. Moeslund. Multimodal and multiview distillation for real-time player detection on a football field. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW), CVsports*, pages 3846–3855, Seattle, WA, USA, June 2020. 3

[16] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 12350 of *Lect. Notes Comput. Sci.*, pages 436–454. Springer, 2020. 2, 4

[17] DeepSORT in PaddleDetection. https://github.com/PaddlePaddle/PaddleDetection/tree/release/2.3/configs/mot/deepsort. 2, 6

[18] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW), CVsports*, pages 4508–4519, Nashville, TN, USA, June 2021. 2, 3, 4

[19] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv*, abs/2003.09003, 2020. 2, 4, 6, 7

[20] Patrick Dendorfer and Heon Song. Motchallenge evaluation kit. https://github.com/dendorferpatrick/MOTChallengeEvalKit, 2020. 6

[21] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 304–311, Miami, FL, USA, June 2009. Inst. Elect. and Electron. Engineers (IEEE). 6

[22] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 1–8, Anchorage, AK, USA, June 2008. Inst. Elect. and Electron. Engineers (IEEE). 6

[23] FairMOT in PaddleDetection. https://github.com/PaddlePaddle/PaddleDetection/tree/release/2.3/configs/mot/fairmot. 6

[24] Baba Fakhar, Hamidreza Rashidy Kanan, and Alireza Behrad. Event detection in soccer videos using unsupervised learning of spatio-temporal features based on pooled spatial pyramid model. *Multimedia Tools and Applicat.*, 78(12):16995–17025, June 2019. 2

[25] Na Feng, Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, Yizhu Zhao, Yunfeng He, and Tao Guan. SSET: a dataset for

shot segmentation, event detection, player tracking in soccer videos. *Multimedia Tools and Applicat.*, 79(39):28971–28992, Oct. 2020. 3, 4, 5

[26] James Ferryman and Ali Shahrokni. PETS2009: Dataset and challenge. In *IEEE Int. Work. Perform. Evaluation Track. Surveill. (PETS)*, pages 1–6, Snowbird, UT, USA, Dec. 2009. 2

[27] Pascual Figueroa, Neucimar Leite, Ricardo ML Barros, Isaac Cohen, and Gerard Medioni. Tracking soccer players using the graph representation. In *IEEE Int. Conf. Pattern Recogn. (ICPR)*, volume 4, pages 787–790, Cambridge, UK, Aug. 2004. IEEE. 3

[28] Xin Gao, Xusheng Liu, Taotao Yang, Guilin Deng, Hao Peng, Qiaosong Zhang, Hai Li, and Junhui Liu. Automatic key moment extraction and highlights generation based on comprehensive soccer video understanding. In *IEEE Int. Conf. Multimedia and Expo Work. (ICMEW)*, pages 1–6, London, UK, July 2020. 2

[29] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv*, abs/2107.08430, 2021. 6

[30] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 3354–3361, Providence, RI, USA, June 2012. 2, 4

[31] Sebastian Gerke, Antje Linnemann, and Karsten Müller. Soccer player recognition using spatial constellation features and jersey number recognition. *Comp. Vis. and Image Underst.*, 159:105–115, June 2017. 3

[32] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW)*, pages 1711–1721, Salt Lake City, UT, USA, June 2018. 2, 3

[33] Samuel Hurault, Coloma Ballester, and Gloria Haro. Self-supervised small soccer player detection and tracking. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 9–18, Seattle, WA, USA, Oct. 2020. 3

[34] Maxime Istasse, Julien Moreau, and Christophe De Vleeschouwer. Associative embedding for team discrimination. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW), CVsports*, pages 2477–2486, Long Beach, CA, USA, June 2019. 3

[35] Sachiko Iwase and Hideo Saito. Parallel tracking of all soccer players by integrating detected positions in multiple view images. In *IEEE Int. Conf. Pattern Recogn. (ICPR)*, volume 4, pages 751–754, Cambridge, UK, Aug. 2004. 3

[36] Haohao Jiang, Yao Lu, and Jing Xue. Automatic soccer video event detection based on a deep neural network combined CNN and RNN. In *IEEE Int. Conf. Tools with Artif. Intell. (ICTAI)*, pages 490–494, San Jose, CA, USA, Nov. 2016. 2

[37] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. SoccerDB: A large-scale database for comprehensive video understanding. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 1–8, 2020. 2

[38] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82(1):35–45, 1960. 6

[39] Muhammad Zeeshan Khan, Summra Saleem, Muhammad A. Hassan, and Muhammad Usman Ghanni Khan. Learning deep C3D features for soccer video event detection. In *Int. Conf. Emerging Technol. (ICET)*, pages 1–6, Islamabad, Pakistan, Nov. 2018. 2

[40] Maria Koshkina, Hemanth Pidaparthy, and James H. Elder. Contrastive learning for sports video: Unsupervised player classification. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW)*, pages 4528–4536, Nashville, TN, USA, June 2021. 3

[41] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv*, abs/1504.01942, 2015. 2

[42] Gen Li, Shikun Xu, Xiang Liu, Lei Li, and Changhu Wang. Jersey number recognition with semi-supervised spatial transformer network. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW)*, pages 1783–1790, Salt Lake City, UT, USA, June 2018. 3

[43] Hengyue Liu and Bir Bhanu. Pose-guided R-CNN for jersey number recognition in sports. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW)*, pages 2457–2466, Long Beach, CA, USA, 2019. 3

[44] Tingxi Liu, Yao Lu, Xiaoyu Lei, Lijing Zhang, Haoyu Wang, Wei Huang, and Zijian Wang. Soccer video event detection using 3D convolutional networks and shot boundary detection via deep feature distance. In *Int. Conf. Neural Inf. Process.*, volume 10635 of *Lect. Notes Comput. Sci.*, pages 440–449. Springer, 2017. 2

[45] Yang Liu, Luiz G. Hafemann, Michael Jamieson, and Mehrsan Javan. Detecting and matching related objects with one proposal multiple predictions. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 4520–4527, Nashville, TN, USA, June 2021. 3

[46] Wei-Lwun Lu, Jo-Anne Ting, James J. Little, and Kevin P. Murphy. Learning to track and identify players from broadcast sports videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1704–1716, July 2013. 3

[47] Jonathon Luiten and Arne Hoffhues. Trackeval. https://github.com/JonathonLuiten/TrackEval, 2020. 6

[48] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *Int. J. Comp. Vis.*, 129(2):548–578, Oct. 2021. 6

[49] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artif. Intell.*, 293, Apr. 2021. 2

[50] Siwei Lyu, Ming-Ching Chang, Dawei Du, Wenbo Li, Yi Wei, Marco Del Coco, Pierluigi Carcagnì, Arne Schumann, Bharti Munjal, Doo-Hyun Choi, et al. UA-DETRAC 2018: Report of AVSS2018 & IWT4S challenge on advanced traffic monitoring. In *IEEE Int. Conf. Adv. Video and Signal Based Surveillance (AVSS)*, pages 1–6, Auckland, New Zealand, Nov. 2018. 2

[51] Mehrtash Manafifard, Hamid Ebadi, and Hamid Abrishami Moghaddam. A survey on player tracking in soccer videos. *Comp. Vis. and Image Underst.*, 159:19–46, June 2017. 3

[52] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv*, abs/1603.00831, 2016. 2, 4, 6

[53] Mahdiar Nekoui, Fidel Omar Tito Cruz, and Li Cheng. FALCONS: FAast Learner-grader for CONtorted poses in Sports. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW)*, pages 900–901, Seattle, WA, USA, June 2020. 3

[54] Peter Nillius, Josephine Sullivan, and Stefan Carlsson. Multi-target tracking-linking identities using Bayesian network inference. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, volume 2, pages 2187–2194, New York City, NY, USA, June 2006. 3

[55] Malte Pedersen, Joakim Bruslund Haurum, Stefan Hein Bengtson, and Thomas B. Moeslund. 3D-ZeF: A 3D zebrafish tracking benchmark dataset. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 2426–2436, Seattle, WA, USA, June 2020. 2, 4

[56] Svein Arne Pettersen, Dag Johansen, Håvard Johansen, Vegard Berg-Johansen, Vamsidhar Reddy Gaddam, Asgeir Mortensen, Ragnar Langseth, Carsten Griwodz, Håkon Kvale Stensland, and Pål Halvorsen. Soccer video and player position dataset. In *ACM Multimedia Syst. Conf.*, Singapore, Singapore, Mar. 2014. Assoc. for Comput. Machinery. 3

[57] Upendra M. Rao and Umesh C. Pati. A novel algorithm for detection of soccer ball and player. In *Int. Conf. Commun. and Signal Process. (ICCSP)*, pages 344–348, Melmaruvathur, India, Apr. 2015. 3

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, June 2017. 3

[59] Francisco Romero-Ferrero, Mattia G. Bergomi, Robert C. Hinz, Francisco J. H. Heras, and Gonzalo G. de Polavieja. idtracker. ai: tracking all individuals in small or large collectives of unmarked animals. *Nat. methods*, 16(2):179–182, Jan. 2019. 2

[60] Melike Şah and Cem Direkoğlu. Evaluation of image representations for player detection in field sports using convolutional neural networks. In *Int. Conf. Theory Appl. Fuzzy Syst. Soft Comput.*, volume 896 of *Adv. in Intell. Syst. and Comput.*, pages 107–115. Springer, 2019. 3

[61] Ryan Sanford, Siavash Gorji, Luiz G. Hafemann, Bahareh Pourbabaee, and Mehrsan Javan. Group activity detection from trajectory and video data in soccer. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW)*, pages 898–899, Seattle, WA, USA, June 2020. 3

[62] Himangi Saraogi, Rahul Anand Sharma, and Vijay Kumar. Event recognition in broadcast soccer videos. In *Indian Conf. Comput. Vision, Graph. Image Process.*, pages 1–7, Dec. 2016. 2

[63] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. CrowdHuman: A benchmark for detecting human in a crowd. *arXiv*, abs/1805.00123, 2018. 7

[64] Mohamad-Hoseyn Sigari, Hamid Soltanian-Zadeh, and Hamid-Reza Pourreza. A framework for dynamic restructuring of semantic video analysis systems based on learning attention control. *Image and Vis. Comp.*, 53:20–34, 2016. Event-based Media Processing and Analysis. 2

[65] Zikai Song, Zhiwen Wan, Wei Yuan, Ying Tang, Junqing Yu, and Yi-Ping Phoebe Chen. Distractor-aware tracker with a domain-special optimized benchmark for soccer player tracking. In *Int. Conf. Multimedia Retrieval*, page 276–284, Taipei, Taiwan, Aug. 2021. Assoc. for Comput. Machinery. 3

[66] Josephine Sullivan and Stefan Carlsson. Tracking and labelling of interacting multiple targets. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 3953 of *Lect. Notes Comput. Sci.*, pages 619–632. Springer, 2006. 3

[67] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):104–119, Jan. 2019. 2

[68] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking pedestrian heads in dense crowd. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 3865–3875, Nashville, TN, USA, June 2021. 2, 4

[69] Inc. SuperAnnotate AI. SuperAnnotate. https://superannotate.com/. 4

[70] Mostafa Tavassolipour, Mahmood Karimian, and Shohreh Kasaei. Event detection and summarization in soccer videos using Bayesian network and copula. *IEEE Trans. Circuits and Syst. for Video Technol.*, 24(2):291–304, Feb. 2014. 2

[71] Rajkumar Theagarajan, Federico Pala, Xiu Zhang, and Bir Bhanu. Soccer: Who has the ball? generating visual analytics and player statistics. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW)*, pages 1749–1757, Salt Lake City, UT, USA 2018. 3

[72] Quang Tran, Bac Vo, Tien Dinh, and Duc Duong. Automatic player detection, tracking and mapping to field model for broadcast soccer videos. In *Int. Conf. Adv. Mob. Comput. Multimedia (MoMM)*, pages 240–243. ACM Press, Dec. 2011. 3

[73] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 12356 of *Lect. Notes Comput. Sci.*, pages 107–122, 2020. 6

[74] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Comp. Vis. and Image Underst.*, 193, Apr. 2020. 2

[75] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. *arXiv*, abs/1604.01850, 2016. 6

[76] Junliang Xing, Haizhou Ai, Liwei Liu, and Shihong Lao. Multiple player tracking in sports video: A dual-mode two-way Bayesian inference approach with progressive observation modeling. *IEEE Trans. Image Process.*, 20(6):1652–1667, June 2010. 3

[77] Yukun Yang, Min Xu, Wanneng Wu, Ruiheng Zhang, and Yu Peng. 3D multiview basketball players detection and localization based on probabilistic occupancy. In *Digit. Image Comp.: Tech. and Applicat.*, pages 1–8, Canberra, ACT, Australia, Dec. 2018. 3

[78] Junqing Yu, Aiping Lei, Zikai Song, Tingting Wang, Hengyou Cai, and Na Feng. Comprehensive dataset of broadcast soccer videos. In *IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, pages 418–423, Miami, FL, USA, June 2018. 3

[79] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. CityPersons: A diverse dataset for pedestrian detection. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 4457–4465, Honolulu, HI, USA, July 2017. Inst. Elect. and Electron. Engineers (IEEE). 6

[80] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. *arXiv*, abs/2110.06864, 2021. 2, 6

[81] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comp. Vis.*, 129:3069–3087, Sept. 2021. 2, 6

[82] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, and Qi Tian. Person re-identification in the wild. *arXiv*, abs/1604.02531, 2016. 6