TeamTrack:
A Dataset for Multi-Sport Multi-Object Tracking in Full-pitch Videos

Atom Scott[1,4,*], Ikuma Uchida[2,4], Ning Ding[1], Rikuhei Umemoto[1], Rory Bunker[1], Ren Kobayashi[1],
Takeshi Koyama[3], Masaki Onishi[4], Yoshinari Kameda[2], Keisuke Fujii[1]

[1]Nagoya University, [2]University of Tsukuba, [3]Tokai University, [4]AIST

*Corresponding Author
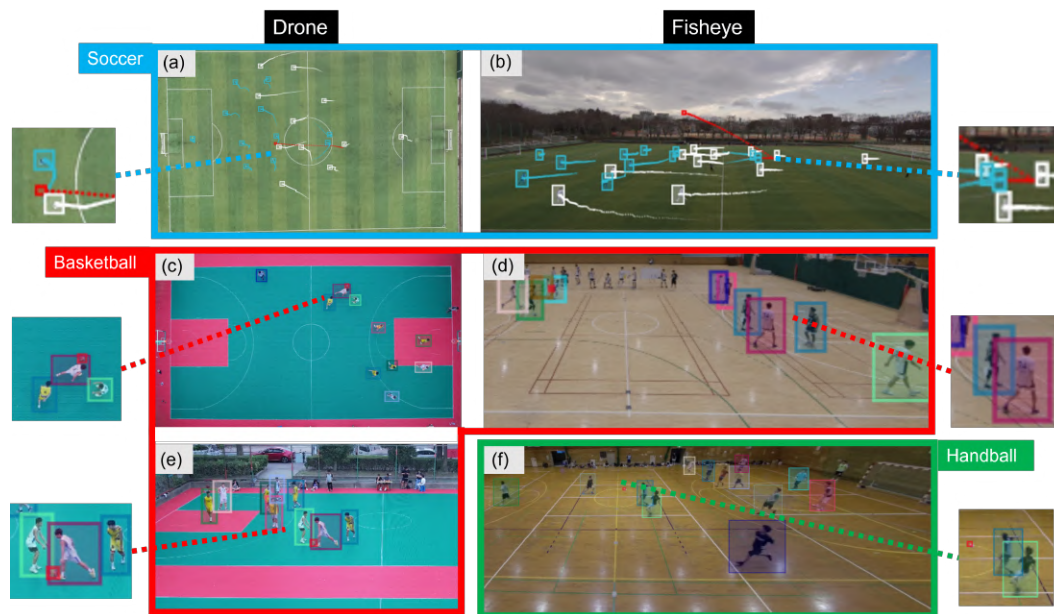
https://atomscott.github.io/TeamTrack/

Figure 1. Example of our TeamTrack dataset. We provide (a) top-view and (b) side-view in soccer, (c) top-view and, (d)/(e) two side-view videos in basketball, and (f) a single side-view video in handball.

Abstract

Multi-object tracking (MOT) is a critical and challenging task in computer vision, particularly in situations involving objects with similar appearances but diverse movements, as seen in team sports. Current methods, largely reliant on object detection and appearance, often fail to track targets in such complex scenarios accurately. This limitation is further exacerbated by the lack of comprehensive and diverse datasets covering the full view of sports pitches. Addressing these issues, we introduce TeamTrack, a pioneering benchmark dataset specifically designed for MOT in sports. TeamTrack is an extensive collection of full-pitch video data from various sports, including soccer, basketball, and handball. Furthermore, we perform a comprehensive analysis and benchmarking effort to underscore TeamTrack's utility and potential impact. Our work signifies a crucial step forward, promising to elevate the precision and effectiveness of MOT in complex, dynamic settings such as team sports. The dataset, project code and competition is released at: https://atomscott.github.io/TeamTrack/.

1. Introduction

Multi-object tracking (MOT) is a key task in computer vision, critical for many practical applications such as autonomous driving and human movement analysis [24, 27,

29, 35]. MOT involves localizing and associating objects over time while dealing with unsolved challenges like occlusion, mis-detections, and ID switches. Current methods typically employ detection followed by re-identification (ReID), relying on appearance-based features for temporal association [40, 45]. However, these methods falter when confronted with objects of similar appearances performing complex movements, which are common in team sports.

In sports analytics, the use of advanced metrics has increased, and the need for fine-grained tracking data has become increasingly evident [7, 31, 36]. Although sensor-based systems, such as the global navigation satellite system and local positioning systems, have emerged as alternatives, these systems have constraints related to availability, budget, and feasibility [15, 26]. Additionally, gaining information on opposing teams can be problematic. As a result, MOT remains a desirable solution, providing a non-intrusive and affordable method for acquiring data.

In light of these considerations, we introduce the Team-Track Dataset, a specialized resource developed to address the unique challenges posed by team sports tracking. Unlike existing datasets, which predominantly feature pedestrian movements or are derived from broadcast footage with limited field coverage, TeamTrack offers an expansive view of the game by incorporating over 200,000 frames and 4 million bounding boxes across football, handball, and basketball, marking it as the largest dataset of its kind in terms of volume and scope, as shown in Table 1. This extensive compilation provides a panoramic view of game scenarios to capture dense player formations, swift motion variations, and frequent occlusions, thereby setting a new benchmark for MOT research in sports.

The primary contributions of this paper are three-fold: (1) We create a new dataset[1], comprising of an unprecedented volume of high-resolution, full-pitch video data from diverse team sports, and multiple viewpoints. (2) A detailed exploration of the dataset's development process, its strategic focus, and an analysis highlighting its characteristics, which promise significant advancements in sports analytics and broader MOT applications. (3) We perform comprehensive evaluations of object detection, trajectory forecasting, and multiple object tracking tasks, setting benchmarks for future research and development.

2. Related Work

Multi-Object Tracking. The typical approach to MOT algorithms follows the tracking-by-detection paradigm. This procedure starts with an object detection using models like RetinaNet [19], CenterNet [9], or YOLO [28], followed by association via feature extraction, typically achieved by using CNNs [13, 45] and recently ViT models [14, 30]. Deep-

---

[1]Available on Kaggle (kaggle.com/datasets/atomscott/ teamtrack) or Google drive (https://bit.ly/3CMN2hP).

Dataset	Frames	BBoxes	Domain
MOT16 [24]	11,235	292,733	Pedestrians
MOT20 [8]	13,410	2,102,385	Pedestrians
KITTI-T [12]	10,870	65,213	Autonomous Driving
DanceTrack [35]	105,855	-	Dance
SSET [10]	12,000	12,000	Soccer
SN-Tracking [5]	225,375	3,645,661	Soccer
SportsMOT [6]	1879	62218	Soccer Basketball Volleyball
SoccerTrack[29]	82,800	2,484,000	Soccer
TeamTrack (ours)	**279,900**	**4,374,900**	**Soccer Basketball Handball**

Table 1. Comparative overview of MOT datasets, showcasing TeamTrack with the highest number of frames and bounding boxes.

SORT [40] exemplifies this approach, combining Kalman motion states and deep appearance descriptors for robust tracking. However, these methods often struggle to distinguish and track objects with similar appearances.

End-to-end tracking methodologies, on the other hand, handle object detection and tracking concurrently, offering potential improvements in performance. Models like Tracktor [2] leverage frame redundancy to eliminate separate data association, while Neural Solver [3] and DeepMOT [41] deploy neural and Siamese networks, respectively, to address tracking. Recent developments have seen the introduction of Transformer architectures into tracking, such as in DETR [4], formulating object detection as a set prediction problem. This paradigm shift has been adopted by models like Trackformer [23] and TransTrack [34]. More advanced models, like MOTR [42] and its improved version MOTRv2 [44], extend this framework by adding a query-interaction module to enhance tracking performance.

Tracking in Sports. In sports tracking, a survey [22] summarized various approaches using a combination of background subtraction, triangulation from multiple cameras, and Kalman filters to track player movements on the pitch [17]. Additionally, research exists which addresses this issue by representing player positions as nodes on a graph, and trajectories as edges between nodes [11, 32]. Sullivan et al. proposed a Bayesian framework for linking player IDs [33]. Furthermore, Lu et al. presented a learning approach based on handcrafted visual features and a Kalman filter to identify and track players within videos [43].

Recent methods have largely focused on using deep learning techniques for player tracking. For instance, Hurault et al. proposed a method of detecting and tracking soccer players using a self-supervised learning method, by transfer learning from an object detection model trained on

generic objects [16]. Theagarajan et al. identified the player holding the ball using a YOLOv2 network detector and a DeepSORT tracker [37]. Maglo et al. utilized human annotations collected in a semi-interactive system [21]. Wang et al. proposed improving tracking in various sports scenes by introducing a three-stage matching process to solve motion blur and body overlap [38].

Multi-Object Tracking Datasets. Development of MOT algorithms depends on the availability of annotated datasets. Early examples are the MOT Challenge dataset [24] and the KITTI Tracking Benchmark [12], focusing on a range of scenarios from crowded urban environments to autonomous driving. Later datasets like UA-DETRAC [39], PETS [25], and DanceTrack [35] further contributed to the field, annotated with bounding boxes and tracking IDs to evaluate algorithmic tracking accuracy.

Tracking in sports presents unique challenges, such as difficulty in differentiating players on the same team and handling occlusion scenarios. While datasets like SoccerNet-Tracking [5] and SoccerTrack [29] provide valuable resources, their focus is often limited to specific sports or partial field views. Furthermore, the use of broadcast footage in datasets such as SportsMOT [6] complicates the direct application of tracking algorithms, necessitating additional processing for image registration, handling zoom, and scene transitions.

The landscape of MOT datasets shows a notable lack of resources tailored for the nuanced challenges of sports tracking. Addressing this, TeamTrack introduces a high-resolution, full-pitch dataset across soccer, basketball, and handball, providing over 279,900 frames and 4,374,900 bounding boxes—the largest in its field. It supports the development of algorithms for complex team sports dynamics. With sports diversity and two perspectives—side and top views—TeamTrack offers comprehensive playfield insights, fostering the development of more sophisticated MOT methods. This dataset aims to drive research forward, making high-quality tracking data accessible and encouraging the creation of robust algorithms to navigate sports tracking's unique challenges.

3. TeamTrack Dataset

In this section, we present TeamTrack, a novel MOT dataset comprising over 150 minutes of high-resolution video from multiple team sports. The main characteristics of TeamTrack are the following:

Large-scale: TeamTrack introduces over 4 million annotated bounding boxes across various tracklets, making it one of the largest datasets of its kind, as compared to others listed in Table 2.

Multiple Sports: TeamTrack includes matches from three team sports; soccer, basketball, and handball.

Similar Appearance and Dynamic Movement: The dataset features targets with similar appearance, dynamic movements, and frequent occlusions, offering a robust challenge for tracking algorithms.

Full Pitch Multi Angle View: Videos are recorded from two angles, top view via drones and side view with fisheye lenses, both covering the entire playing field.

Beyond its significant size and the inherent challenges it presents, TeamTrack's unique value lies in its multi-view and full-pitch capture settings. These features enable experimentation with multi-view tracking and the use of prior information (e.g., player count, pitch dimensions) not feasible with traditional broadcast videos. Additionally, we hope the adversarial-cooperative nature of team sports can also be studied to further refine tracking techniques.

In the following subsections, we describe the data collection process, annotation and labeling, appearance similarity, and motion patterns observed in the TeamTrack dataset, following the methodology described in DanceTrack [35]. Our dataset was compared to DanceTrack, MOT17, and MOT20—datasets characterized by limited camera movements. We did not compare with broadcast video datasets, such as SoccerNet-Tracking [5] and SportsMOT [6], due to their extensive camera movements, which hinder fair comparison.

3.1. Data Collection

The data collection process captured video footage across three team sports: soccer, basketball, and handball, recorded at various university venues, encompassing both indoor and outdoor environments. The side and top views were recorded using fisheye and drone cameras, respectively. To correct for the distortion caused by the fisheye lens, we applied Zhang's calibration method. Further information on the devices used and their resolution specifications is detailed in Table 2.

3.2. Annotation and Labeling

The annotation process involved a series of strategic decisions and tool evaluations to balance efficiency, accuracy, and cost. Initially, we utilized the open-source tool CVAT but later transitioned to Labelbox to alleviate the burdens associated with working on a local server and to streamline workflows more effectively. Although Labelbox did also offer automated tracking, the practical gains in speed were limited due to inference time not being fast enough.

Interpolation and Manual Annotation: Both CVAT and Labelbox support interpolation, significantly reducing the need to annotate every frame. This feature enabled focused manual annotation on approximately every 1 to 50 frames, as depicted in the timeline of Fig. 2, especially in scenes experiencing dynamic changes. While this method streamlined the workflow, it still required significant manual input to ensure accuracy.

Table 2. Camera and video details of the multi-object tracking dataset for various sports.

Sport	Perspective	Location	Device	Resolution	Minutes	Bounding Box Count
Soccer	Side (Fisheye)	University of Tsukuba, Outdoor	Z CAM E2-F8	8K	30	1,242,000
Soccer	Top (Drone)	University of Tsukuba, Outdoor	DJI Mavic 3	4K	30	1,242,000
Basketball	Side (Fisheye)	Tokai University, Indoor	Z CAM E2-F6	6K	17.5	346,500
Basketball	Top (Drone)	Wuhu Institute of Technology, Outdoor	DJI Mavic 3	4K	24	475,200
Basketball	Side (Drone)	Wuhu Institute of Technology, Outdoor	DJI Mavic 3	4K	24	475,200
Handball	Side (Fisheye)	Nagoya University, Indoor	Z CAM E2-F6	6K	30	594,000
Total					155.5	4,374,900

Figure 2. Screenshot of Labelbox during the annotation process, illustrating manual adjustments to automated tracking and interpolation annotations.
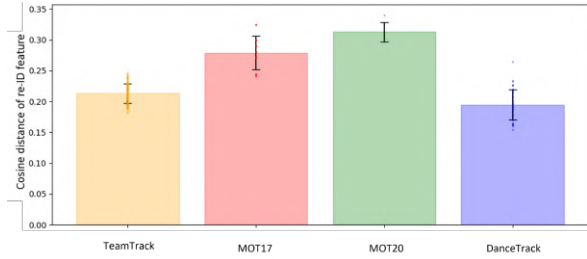

Figure 3. Cosine distance of re-ID features: TeamTrack compared to DanceTrack, MOT17, and MOT20.
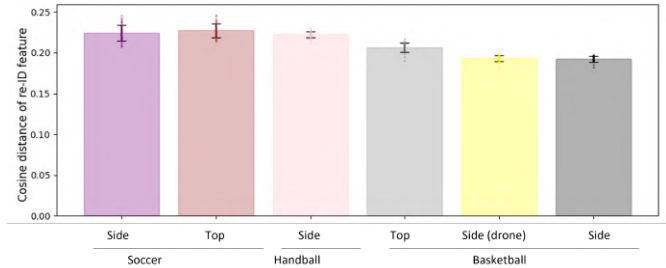
Figure 4. Cosine distance of re-ID features within TeamTrack datasets.

Challenges with Semi-Automatic Detection: We experimented with semi-automatic detection using a pretrained object detector for initial annotations. However, the time saved was often offset by the need to correct false positives/negatives and manage ID switches. Labelbox's lack of track merging capabilities further complicated these adjustments, as illustrated by the comparative analysis of pretrained versus fine-tuned object detectors' accuracy in Table 3, indicating the necessity for model fine-tuning.

Exploration of Other Tools: In search of more efficient solutions, we considered advanced annotation tools like V7 and Encord, known for their sophisticated track merging and smart interpolation capabilities. However, the high initial limited their adoption.

Annotation Effort: Ultimately, the annotation process required over 600 person-hours, translating to roughly 2 hours for every 30 seconds of video. This significant investment of time and resources highlighted a critical challenge: the pressing need for more efficient annotation tools to facilitate dataset expansion and enhance the scalability of sports analytics research. We hope these insights offer valuable perspectives to the research community on balancing the capabilities, efficiency, and costs of annotation tools.

Dataset Metadata: The TeamTrack dataset provides persistent player IDs throughout the entirety of a match, although it does not include jersey numbers. Annotations are available in two formats: the SportsLabKit format[2], which encompasses team affiliations to enrich analyses of team dynamics and interactions, and the standard MOTChallenge format[3]. The latter is especially useful for general tracking studies and conforms to conventional evaluation methodologies.

3.3. Appearance Similarity

Quantitave analysis of appearance features: To quantitatively test our hypothesis that the targets in teamtrack have similar appearances, we employ a pre-trained re-ID model [35] to extract the appearance features and calculate a similarity metric. The similarity metric is based on the cosine

---

[2] https://github.com/AtomScott/SportsLabKit.
[3] https://motchallenge.net/instructions/.

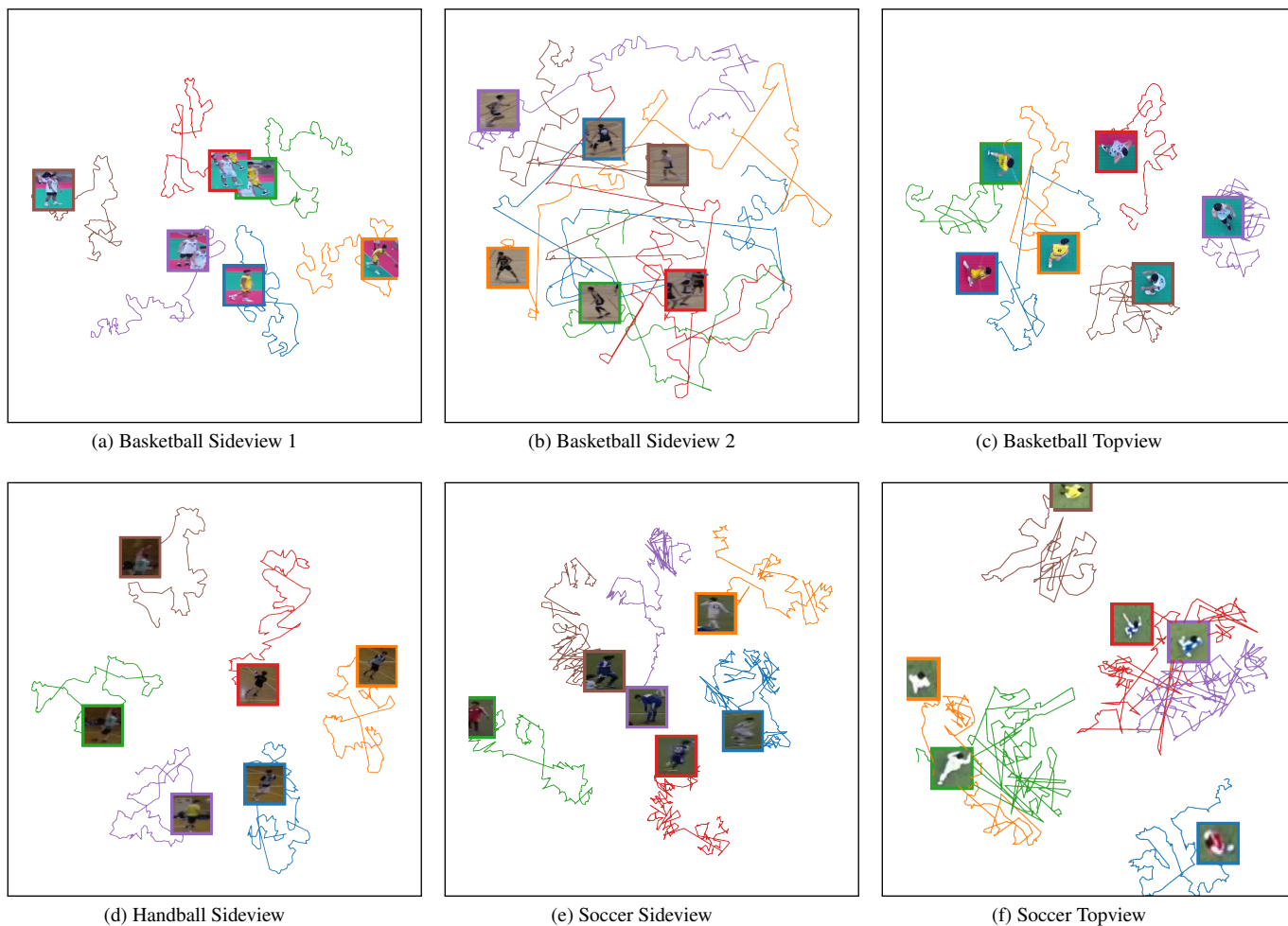| (a) Basketball Sideview 1	(b) Basketball Sideview 2	(c) Basketball Topview
(d) Handball Sideview	(e) Soccer Sideview	(f) Soccer Topview

Figure 5. Visualization of re-ID features from various videos in our TeamTrack dataset using t-SNE. Objects are color-coded for consistent identification. The bounding box from the first frame is superimposed on the corresponding re-ID feature for contextualization.

distances between these features across video frames:

$$V = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{N_t^2} \sum_{i=1}^{N_t} \sum_{j \neq i}^{N_t} \left(1 - \cos < F(B_{ti}), F(B_{tj}) >\right),$$

where $F(B_{ti})$ and $F(B_{tj})$ are appearance features, B and T are an object and the number of frames in the video, N_t is the number of objects on the frame t and $< \cdot >$ is the angle between two vectors. As depicted in Figure 3, TeamTrack datasets generally demonstrated a higher appearance similarity among players compared to the MOT datasets, with lower mean cosine distances. Figure 4 shows the comparison within TeamTrack datasets, highlighting that the similarity metric is somewhat consistent across sports. **Visualization of appearance features**: To further understand the visual characteristics of the TeamTrack dataset, we present visualizations of appearance features extracted from several videos. In Figure 5, we show t-SNE embeddings of image Re-ID features, extracted from the first 200

frames of each dataset. These visualizations reveal that, although appearance features can be differentiated in certain sports—particularly under varying background conditions and among players from opposing teams—features of players within the same team tend to be highly entangled. It should be noted that as the number of frames increases, the bounding boxes for players are likely to overlap, adding another layer of complexity to the task. This qualitatively demonstrates the high degree of appearance similarity among within the proposed TeamTrack dataset.

3.4. Motion Patterns

TeamTrack dataset includes complex group motion patterns compared to popular pedestrian datasets. To quantitatively confirm this, we computed IoU (Intersection-over-Union) on adjacent frames and frequency of relative position switch as motion patterns metrics in the TeamTrack dataset and compare them with other MOT datasets, as described in DanceTrack [35].
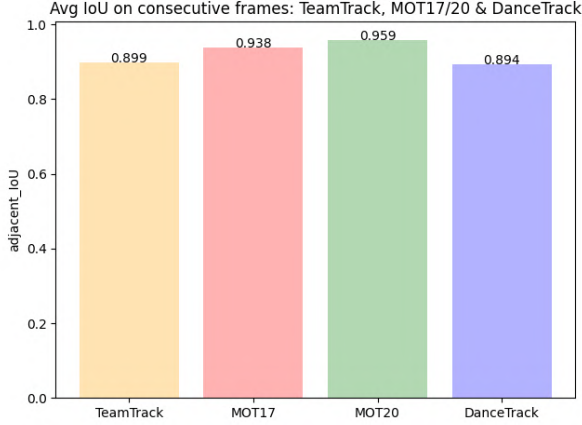
Figure 6. Average IoU on adjacent frames: TeamTrack compared with MOT17, MOT20, and DanceTrack.
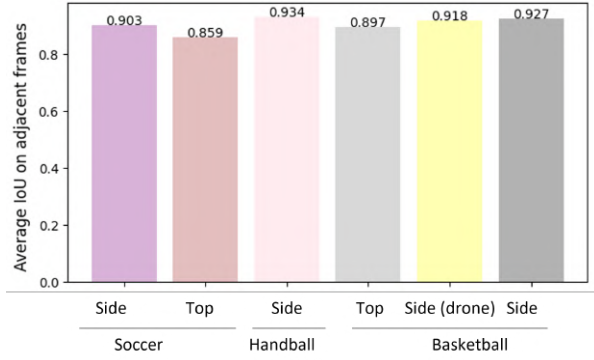

Figure 7. Average IoU on adjacent frames: within TeamTrack datasets.

IoU on adjacent frames: We calculate the average IoU of bounding boxes on adjacent frames for each video. A low IoU implies fast-moving objects or low video frame rates. The averaged IoU on adjacent frames for a video with N objects and T frames is defined as follows:

$$U = \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T-1} IoU(B_{ti}, B_{t+1i}).$$

Figure 6 shows the IoU on adjacent frames in TeamTrack and other datasets. The results indicate that TeamTrack had a slightly lower average IoU on adjacent frames compared to MOT17/MOT20, and a similar average IoU on adjacent frames to DanceTrack. The lower average IoU on adjacent frames suggests that objects/regions move, change in shape or size, or disappear to a greater extent in TeamTrack and DanceTrack compared to MOT17/MOT20.

Figure 7 shows the results among the six TeamTrack datasets. Soccer Top View and Handball Side View had the lowest and highest average IoU on adjacent frames, respectively, indicating that objects maintained their position, size, and shape to a greater extent in handball.
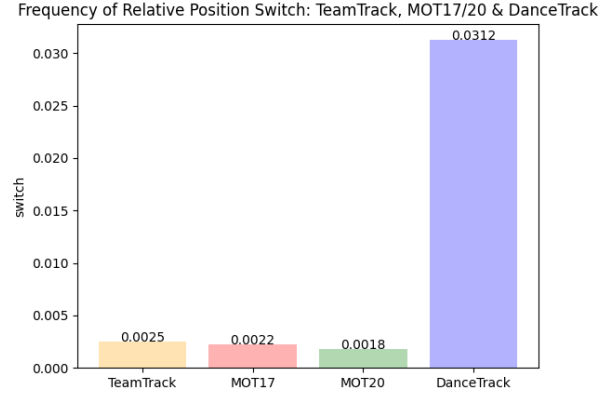

Figure 8. Frequency of relative position switch: TeamTrack compared with MOT17, MOT20, and DanceTrack
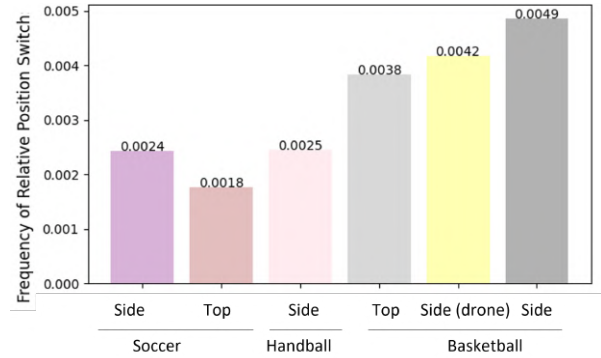

Figure 9. Frequency of relative position switch: within TeamTrack datasets.

Frequency of relative position switch: To measure the diversity of object motions, we compute the average frequency of relative position switches for each video based on [35]. Frequent position switches can be attributed to highly non-linear motion patterns and result in frequent crossovers and inter-object occlusions. The average frequency of relative position switch is defined as follows:

$$S = \frac{\sum_{i=1}^{N} \sum_{j \neq i}^{N} \sum_{t=1}^{T-1} sw(B_{ti}, B_{tj}, B_{t+1i}, B_{t+1j})}{2N(T-1)(N-1)},$$

where sw is an indicator function, where $sw(\cdot) = 1$ if the two objects swap their left-right relative position or top-down relative position on the adjacent frames, $sw(\cdot) = 0$ if there is no swap.

Figure 8 shows the results in TeamTrack and other datasets, with DanceTrack showing significantly higher frequencies and TeamTrack marginally exceeding those of MOT datasets. This likely stems from the dynamics of team sports, where intense activity often concentrates around the vicinity of the ball or specific events such as corner kicks, potentially leading to an underrepresentation in averaged

data. Further analysis within TeamTrack, as illustrated in Figure 9, reveals considerable differences among sports. In the context of sports, we interpret the frequency of relative position switch to indicate the activity/intensity level of the match; therefore, in this case, Basketball had a higher level of intensity than Soccer and Handball.

4. Experiments and Evaluation

Our evaluation showcases the utility of the TeamTrack dataset and the distinct challenges it introduces for advancing research in MOT. We concentrate on three pivotal components: object detection (Section 4.1), trajectory forecasting (Section 4.2), and finally, multi-object tracking (Section 4.3). In each task we followed a conventional data splitting ratio of 70:15:15 for training, validation, and testing sets.

4.1. Object Detection

Object detection is a crucial step for accurately tracking an object. Detection accuracy typically correlates with tracking performance, particularly with methods that employ the tracking-by-detection paradigm.

Setup: We fine-tuned the YOLOv8 model [18] on each dataset. We adjusted the image size using YOLOv8m, YOLOv8n, and YOLOv8x. The image sizes employed were 512, 1280, and 2560. The training utilized the AdamW optimizer[20] with an initial learning rate of 0.001, alongside data augmentation techniques (mosaic, flipping, scaling, color jittering) to enhance model robustness. Test-Time Augmentation (TTA) was also implemented.

Metrics: We evaluate the performance of the model using the mean Average Precision (mAP) with intersection over union (IoU) thresholds ranging from 0.5 to 0.95 ($mAP_{50:95}$). This metric measures the model's accuracy in terms of both position and size of the predicted bounding boxes.

Results: Table 3 showcases YOLOv8's object detection performance across various sports, separated by camera perspectives. The notably lower mAP scores for pre-trained and fine-tuned models in top-view setup can be largely attributed to the smaller appearance of players from high-altitude drone footage. This results in smaller IoU values, making it challenging to achieve high ($mAP_{50:95}$) scores. Nevertheless, fine-tuning significantly improved the mAP scores for all sports and views. The result underscores the importance of fine-tuning for enhancing model performance and highlight the influence of camera perspective on detection accuracy.

4.2. Trajectory Forecasting

Trajectory forecasting involves predicting future positions of subjects and can enhance object tracking. For instance, it facilitates the re-identification or continued tracking of objects lost due to occlusions, This is particularly critical in scenarios where appearance cues are unreliable, such as

Table 3. COCO $mAP_{50:95}$ performance of the object detection model on different sports, split by TV (Top-View) and SV (Side-View) camera perspectives. For basketball, SV* denotes the side view with drone camera.

YOLOv8	Football		Basketball			Handball
	TV	SV	TV	SV	SV*	SV
Pre-trained	1.4	11.2	2.5	15.4	10.5	9.8
Fine-tuned	**23.5**	**52.7**	**66.6**	**68.7**	**68.6**	**71.0**

in team sports. We implement both Constant Velocity and LSTM motion models to analyze the movement patterns of individual players within the teamstrack dataset, aiming to grasp the motion characteristics of the TeamTrack dataset.

Setup: We adopted the experiment settings from social-LSTM [1], observing trajectories for 3.2 seconds and predicting the trajectories for the next 4.8 seconds (144 frames). The data was centralized and normalized, and a 5% overlap sliding window was employed. Data augmentation was carried out by horizontal and vertical flipping on each sequence. Standard procedures, such as teacher forcing and learning rate scheduling, were implemented.

Metrics: Our trajectory forecasting model is assessed based on its Root Mean Square Error (RMSE) for one step ahead prediction. RMSE is a standard metric used in trajectory forecasting as it provides an aggregate measure of prediction error. It quantifies how much, on average, our model's predictions deviate from the actual data. A lower RMSE indicates a more accurate model, providing a reliable means of comparing different models or setups.

Results: Table 4 showcases the RMSE of both motion models across different sports, measured immediately (1 frame) and a few seconds (144 frames) post-observation. Errors tend to accumalte over time. LSTM generally outperforms, validating the hypothesis that TeamTrack dataset has diverse and non-linear movement. Incorporating team labels and models that capture player interactions, such as graph neural networks, could further improve accuracy, presenting an avenue for future research.

Table 4. RMSEs of the trajectory forecasting models on different sports, which are averaged over all trajectories.

Video	Football		Basketball		Handball	
Model/Frames	1	144	1	144	1	144
Constant	0.04	7.99	0.06	7.68	0.04	6.08
LSTM	**0.02**	**6.50**	**0.04**	**5.19**	**0.02**	**4.07**

4.3. Multiple Object Tracking

In this section, we present a direct evaluation our proposed TeamTrack dataset by applying two state-of-the-art tracking algorithms, ByteTrack and BoT-SORT. Additionally, we compare these results with earlier benchmarks performed on the MOT17 and DanceTrack datasets to highlight the challenges posed by our TeamTrack dataset.

Setup: Similar our prior experiments, we split the dataset and proceeded to train YOLOv8 detectors on the training set, followed by hyperparameter optimization using the validation set. For our evaluations, we employed the Ultralytics versions of ByteTrack and BoT-SORT. When comparing to the DanceTrack and MOT17 datasets, we concentrated on ByteTrack alone due to the public availability of results for both datasets, allowing for a direct comparison of performance across all sports within the TeamTrack dataset.

Metrics: To evaluate the different aspects of tracking, we report results using several commonly used metrics: HOTA, MOTA, IDF1, DetA, and AssA. Each metric uniquely emphasizes aspects of detection (DetA), association performance (AssA, IDF1), or both (HOTA, MOTA), providing a comprehensive evaluation of tracking capabilities.

Results: The results across different sports and view perspectives, detailed in Table 5. Both methods showed similar trends. In the Handball SideView, models achieved the highest HOTA, indicating effective tracking capabilities in this setting. The Basketball scenarios, SideView, SideView2 and TopView, revealed contrasting outcomes: the SideView2 posed significant challenges, while the TopView showed better performance metrics, suggesting variations in model efficacy based on the angle of view. In the Soccer datasets, the SideView perspective generally yielded better performance compared to the TopView. Table 6 shows a comparison of these scores compared to those of other datasets. Although TeamTrack's average scores across all sports do not position it as the most challenging dataset, the variability within specific sports and perspectives underscores its diversity and complexity. This variability marks TeamTrack as a uniquely challenging dataset for developing and evaluating tracking algorithms.

5. Conclusion

In this study, we introduced TeamTrack, a dataset for multi-object tracking (MOT). The TeamTrack dataset captures object appearances and movements across football, basketball, and handball games using full-pitch, high-resolution videos. Our work included experiments in object detection, trajectory forecasting, and MOT. Our findings reveal opportunities for further research. Specifically, the TeamTrack dataset has limitations regarding team and venue diversity. Expanding the dataset to include a wider range of teams and locations could improve the model generalization. We hope the TeamTrack dataset will contribute to the development of more effective tracking models.

Table 5. Performance of the various MOT models.

(a) Soccer SideView

Method	HOTA	DetA	AssA	MOTA	IDF1
BoT-SORT	58.4	62.8	54.5	84.2	73.8
ByteTrack	59.3	64.4	54.7	86.4	74.2

(b) Soccer TopView

Method	HOTA	DetA	AssA	MOTA	IDF1
BoT-SORT	51.9	51.1	53.3	42.7	65.7
ByteTrack	53.7	51.4	56.5	43.3	69.2

(c) Basketball SideView

Method	HOTA	DetA	AssA	MOTA	IDF1
BoT-SORT	75.2	79.2	71.4	94.3	85.9
ByteTrack	76.2	75.5	76.9	89.3	88.6

(d) Basketball SideView2

Methods	HOTA	DetA	AssA	MOTA	IDF1
BoT-SORT	47.3	67.6	33.1	80.2	50.8
ByteTrack	42.9	54.7	33.7	65.0	53.6

(e) Basketball TopView

Method	HOTA	DetA	AssA	MOTA	IDF1
BoT-SORT	66.3	62.7	70.3	89.0	93.9
ByteTrack	65.7	65.1	66.4	89.6	92.0

(f) Handball SideView

Methods	HOTA	DetA	AssA	MOTA	IDF1
BoT-SORT	75.1	75.5	74.7	91.6	89.7
ByteTrack	73.5	73.8	73.2	89.4	87.6

Table 6. ByteTrack tracking metrics compared across various datasets, including average scores from TeamTrack and individual scores from sports with the highest (Basketball-SV) and lowest (Basketball-SV2) HOTA scores.

Dataset	HOTA	DetA	AssA	MOTA	IDF1
TeamTrack Ave.	61.9	64.2	60.2	77.2	77.5
Basketball-SV	76.2	75.5	76.9	89.3	88.6
Basketball-SV2	42.9	54.7	33.7	65	53.6
MOT17	63.1	64.5	62	80.3	77.3
DanceTrack	47.1	70.5	31.5	88.2	51.9

6. Acknowledgements

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 7

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 2

[3] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6247–6257, 2020. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2

[5] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3502, 2022. 2, 3

[6] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *arXiv preprint arXiv:2304.05170*, 2023. 2, 3

[7] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *KDD*, pages 1851–1861, 2019. 2

[8] 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 preprint arXiv:2003.09003*, 2020. 2

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 2

[10] 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 Applications*, 79:28971–28992, 2020. 2

[11] Pascual J Figueroa, Neucimar J Leite, and Ricardo ML Barros. Tracking soccer players aiming their kinematical motion analysis. *Computer Vision and Image Understanding*, 101 (2):122–135, 2006. 2

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. The kitti vision benchmark suite. *URL http://www. cvlibs. net/datasets/kitti*, 2(5), 2015. 2, 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[14] Shuting He, Haowen Luo, Pichao Wang, F. Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14993–15002, 2021. 2

[15] Liam Hennessy and Ian Jeffreys. The current use of gps, its potential, and limitations in soccer. *Strength & Conditioning Journal*, 40(3):83–94, 2018. 2

[16] Samuel Hurault, Coloma Ballester, and Gloria Haro. Self-supervised small soccer player detection and tracking. In *Proceedings of the 3rd international workshop on multimedia content analysis in sports*, pages 9–18, 2020. 3

[17] Sachiko Iwase and Hideo Saito. Parallel tracking of all soccer players by integrating detected positions in multiple view images. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, pages 751–754. IEEE, 2004. 2

[18] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 2023. 7

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[21] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Efficient tracking of team sport players with few game-specific annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3461–3471, 2022. 3

[22] Mehrtash Manafifard, Hamid Ebadi, and H Abrishami Moghaddam. A survey on player tracking in soccer videos. *Computer Vision and Image Understanding*, 159:19–46, 2017. 2

[23] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 2

[24] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 2, 3

[25] Luis Patino, Tom Cane, Alain Vallee, and James Ferryman. Pets 2016: Dataset and challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016. 3

[26] 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 *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 18–23, 2014. 2

[27] Akshay Rangesh and Mohan Manubhai Trivedi. 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. 1

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[29] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. Soccertrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3569–3579, 2022. 2, 3

[30] Yao Shen, Lei Wang, and Yue Jin. Aaformer: A multi-modal transformer network for aerial agricultural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1705–1711, 2022. 2

[31] William Spearman. Beyond expected goals. In *Proceedings of the 12th MIT sloan sports analytics conference*, pages 1–17, 2018. 2

[32] Josephine Sullivan and Stefan Carlsson. Tracking and labelling of interacting multiple targets. In *European conference on computer vision*, pages 619–632. Springer, 2006. 2

[33] Josephine Sullivan, Peter Nillius, and Stefan Carlsson. Multi-target tracking on a large scale: Experiences from football player tracking. In *Proceedings of the IEEE ICRA 2009 workshop on people detection and tracking*. Citeseer, 2009. 2

[34] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2

[35] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022. 2, 3, 4, 5, 6

[36] Masakiyo Teranishi, Kazushi Tsutsui, Kazuya Takeda, and Keisuke Fujii. Evaluation of creating scoring opportunities for teammates in soccer via trajectory prediction. In *International Workshop on Machine Learning and Data Mining for Sports Analytics (MLSA)*. Springer, 2022. 2

[37] Rajkumar Theagarajan, Federico Pala, Xiu Zhang, and Bir Bhanu. Soccer: Who has the ball? generating visual analytics and player statistics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1749–1757, 2018. 3

[38] Jie Wang, Yuzhou Peng, Xiaodong Yang, Ting Wang, and Yanming Zhang. Sportstrack: An innovative method for tracking athletes in sports scenes. *arXiv preprint arXiv:2211.07173*, 2022. 3

[39] 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. *Computer Vision and Image Understanding*, 193:102907, 2020. 3

[40] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2

[41] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixé, and Xavier Alameda-Pineda. How to train your deep multi-object tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6787–6796, 2020. 2

[42] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 659–675. Springer, 2022. 2

[43] Yi Zhang, Hanqing Lu, and Changsheng Xu. Collaborate ball and player trajectory extraction in broadcast soccer video. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008. 2

[44] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22056–22065, 2023. 2

[45] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019. 2