SoccerNet 2023 Tracking Challenge - MOT4MOT Team Technical Report

Gal Shitrit*, Ishay Be'ery*, Ido Yerhushalmy
Amazon Prime Video Sports
galshi, ishaybee, idoy@amazon.com

Abstract

The SoccerNet 2023 tracking challenge requires the detection and tracking of soccer players and the ball. In this technical report, we present our approach to tackle these tasks separately. For player tracking, we employ a state-of-the-art online multi-object tracker along with a contemporary object detector. To overcome the limitations of the online approach, we incorporate a post-processing stage that includes interpolation and appearance free track merging. Additionally, an appearance-based track merging technique is used to handle track termination and creation far from the image boundaries. For ball tracking, we treat it as a single object detection problem and utilize a fine-tuned YOLOv8l detector with filtering techniques to enhance detection precision. Our final player and ball tracker achieves competitive results on the SoccerNet 2023 tracking challenge with a HOTA score of 66.27.

1. Introduction

The SoccerNet 2023 tracking challenge presents a unique and challenging task of detecting and tracking soccer players and the ball. Tracking multiple objects in a dynamic and fast-paced sport like soccer is a challenging problem that requires advanced algorithms and techniques. In this technical report, we present our approach to tackling this challenge by treating player tracking and ball tracking as separate tasks.

2. Related Work

Multi-Object Tracking (MOT) is the task of identifying and maintaining multiple object trajectories or tracks from a video stream. For the online scenario, often applied to live video streams, predictions cannot rely on future frames. A common methodology to address this task, is Tracking-by-detection [1–3], a two-step approach that first detects the objects in each frame of the video, and associates these detections over time to form trajectories (tracks). One way to create coherent trajectories is by applying a constant velocity Kalman Filter (KF) [2]. To better distinguish between objects, OC-SORT [4] added a velocity consistency term whereas DeepSORT [1] addressed this gap by integrating additional appearance information [1]. StrongSORT [3] improved the latter approach by replacing the appearance model with Bag of Tricks [5] (BoT), and the addition of a Camera Motion Compensation (CMC). DeepOC-SORT [6] combined components from both StrongSORT and OC-SORT in an holistic manner to benefit from the two approaches. When information from future frames is available, for instance, when there is a time gap between broadcasting and acquiring the video feed, additional offline methods may be applied. StrongSORT++ [3] proposed to impute missing detections by interpolating and smoothing with a Gaussian process (GSI). Additionally, tracks are merged using the Appearance Free Linking (AFLink) model.

3. Method

In soccer games, there is a notable differences between ball tracking and player tracking. Notably, the presence of multiple players in each frame contrasts with the single occurrence of a ball. Additionally, the ball can undergo considerable acceleration within brief time intervals. Furthermore, ball detection poses heightened challenges due to factors such as its relatively small size, frequent occlusions, and tendency to blend with the players' uniforms or crowd. Hence, in light of these considerations, the tracking algorithmic solution was split into distinct components, with separate independent mechanisms designated for ball tracking and player tracking. The overall algorithm flow can be seen in figure 1.

Player Tracking Player tracking was approached through a two-step methodology. In the initial step, an optimized state-of-the-art (SoTA) online multi-object tracker was employed in conjunction with a contemporary object detector. However, due to the inherent limitations of this online approach, which lacks the ability to anticipate future events or modify past outcomes, a post-processing stage was incorporated to refine the tracking data. The post-processing phase comprises three distinct methods. Firstly, missing track de-

*Equal contribution

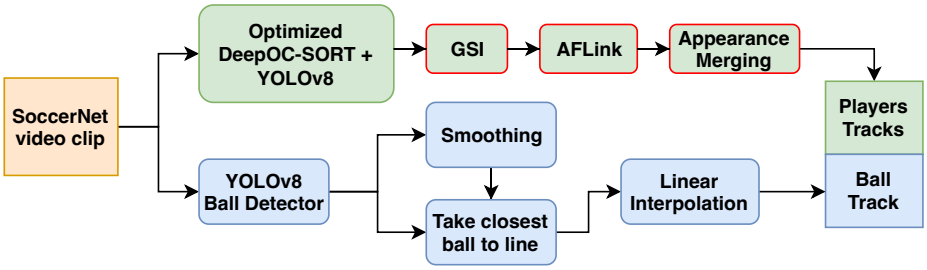

Figure 1. Tracking algorithm flow. The ball tracking components are marked in blue and the player tracking components are marked in green. The post-processing steps for the player tracker are marked with a red line.

tections are interpolated using GSI. Secondly, a fine-tuned AFLink model is utilized to merge tracks. The combination of these two techniques is denoted as "++" in this study. Thirdly, an appearance-based track merging technique is employed. This method operates under the assumption that, in our particular setup, new track IDs can only be created or terminated near the boundaries of the image. As a result, an iterative merging process is employed to identify track IDs that were terminated far from the image boundary. Subsequently, if a new track is generated within a short time span, an attempt is made to merge it with the previously terminated track if their appearance similarity remains consistently high throughout most of the track duration.

Ball Tracking We approached ball tracking by treating it as a single object detection problem. In order to maximize the detection recall, we employed a fine-tuned YOLOv8l [7] detector with a low confidence score of 0.05. This configuration yields multiple detections for each frame within the video clip. To enhance the precision of the detections, we leveraged the temporal nature of the data through a series of filtering techniques. First, the center coordinates of the most confident detection in each frame was smoothed using 3rd order polynomial with a window size of 51 frames. Subsequently, we retained only the detection closest to this smoothed trajectory if its distance was below 100 pixels. To address any missing detections, the resulting detections were then linearly interpolated to ensure tracking box continuity.

4. Experiments

We present a process aimed to improve and assess the performance of the MOT tracker by optimizing its individual components, namely the appearance model, detector and tracker type, followed by their integration and evaluation. The optimization process involved fine-tuning each component in isolation and assessing their effectiveness. Subsequently, the optimized components were integrated into the tracker to evaluate the overall MOT performance.

4.1. Data Preparation and Setup

Player Detector Our tracker is heavily dependent on the performance of the player detector. To improve performance on target domain, a YOLOv8l detection model was fine-tuned to detect players on SoccerNet data and the Average Precision (AP) at IoU of 0.5 was measured. The image resolution was 576x1024, with Non Max Suppression (NMS) IoU threshold of 0.45. The bounding boxes of players were extracted from a total of 1710 frames belonging to the tracking training set. These frames were selected by sampling at a rate of 1 frame per second (FPS) from all available training clips. Additionally, 200 manually inspected frames from the tracking test set were selected for validation. It is important to note that the labels were not modified or changed in any manner during this process.

Ball Detector A separate YOLOv8l ball detection model was fine-tuned to detect the ball on SoccerNet data and the AP at IoU of 0.5 was measured. The image resolution was 1080x1920. The bounding boxes of balls were extracted from a total of 2084 frames belonging to the tracking training set. Due to errors in the ground-truth labels, a COCO2017 pre-trained detector was used to select only frames with IoU$>$0.7 between the label and the ball detection. This procedure ensures that only high quality labels appear in the training set. Additionally, 200 manually inspected frames from the tracking test set were selected for validation. It is important to note that the labels were not modified or changed in any manner during this process.

Appearance Model An appearance model (OSNet-ain [8]) was fine-tuned to match the target domain. Three different sizes of the model were trained the smallest being $\times 0.25$ and the largest $\times 1$. The crop resolution was 256x128. The Rank-1 accuracy was measured along with the mAP. The appearance dataset contained IDs extracted from the tracking train set. To ensure ID consistency between different clips, the team metadata, track ID and jersey number metadata was utilized. Tracks without known jersey number were discarded from the dataset. The dataset comprises six distinct games, encompassing three games for training set

Table 1. Performance of the player detection model on the test set when trained over different datasets.

Training Data	AP@0.5 ↑	AP@0.5:0.95 ↑
COCO2017	0.954	0.812
SoccerNet	**0.990**	**0.923**

and three games for validation set, featuring a total of 199 unique IDs, with 123 IDs in the training set and 76 IDs in the validation set. Notably, the training set encompasses an average of 1600 images per ID, while the validation set is composed of 10 images per ID, featuring eight gallery and two query images, totaling 760 images. The validation images were drawn from the tracking test set.

Player Tracker Two different trackers were evaluated with the fine-tuned components. The appearance model was used with cosine similarity. Furthermore, the upper limit of the tracker's performance was assessed by conducting the same experiment using GT boxes. The HOTA metric [9] was used for evaluation on SoccerNet test set.

5. Results

Player Detector Fine-tuning the player detector on the SoccerNet data improves its performance (AP@0.5 0.954 to 0.994), as can be seen in table 1. After analyzing the failure cases, we found that some errors occur when several players overlap with each other.

Ball Detector Fine-tuning the ball detector detector results in AP@0.5 of 0.95 and AP@0.5:0.95 of 0.71. This indicates that the bounding box produced by the detector is not tight enough. Furthermore, our investigation revealed that the detector struggles in scenarios involving partial occlusion, as well as when the ball's visual characteristics merge with other objects like the crowd or white shoes.

Appearance Model The different model sizes achieved similar mAP scores, with crop augmentation providing the greatest improvement, as can be seen in table 2. The comparability of performance can be attributed to a multitude of factors, including a relatively modest number of identities (199 IDs as opposed to 1501 IDs in Market1501 [10]), which can lead to rapid overfitting, errors in training data that impede model improvement, and nearly indistinguishable appearances of certain players on the same team.

Player Tracker DeepOC-SORT++ achieves slightly better HOTA metric than Strong-SORT++ (66.0% and 65.2% respectively), as can be seen in table 3. The post processing appearance merging further boosts the HOTA by +0.38%. Using GT boxes instead of detections greatly improves the HOTA metric by a large margin of 21.85 points. The results suggest that the efficacy of tracking is closely linked to the precision of the detector. This observation is further supported by the ablation study (see table 4), which demonstrates that the fine-tuned detector has the most substantial impact on the HOTA metric, resulting in a gain of 5.9 points.

Final Tracker The final player and ball tracker achieved HOTA of 66.27, DetA of 70.32 and AssA of 62.62 on the challenge set.

Table 2. Appearance model performance for different model sizes and train augmentations.

Model	Augmentations	Rank-1 ↑	mAP ↑
OSNet-x0.25	None	0.83	0.79
OSNet-x0.25	Crop	0.93	0.77
OSNet-x0.75	Crop	0.95	0.8
OSNet-x0.75	Crop, Flip	**0.95**	**0.8**
OSNet-x1	Crop	0.94	0.78

Table 3. Player tracking performance on SoccerNet test set for different trackers. Post processing appearance merging is denoted by p.

Tracker	Detector	Configuration	HOTA ↑
Strong-SORT++	YOLOv8	OSNet-x1	65.18
DeepOC-SORT++	YOLOv8	OSNet-x0.75	66.00
DeepOC-SORT++p	YOLOv8	OSNet-x0.75	**66.38**
Strong-SORT++	GT boxes	OSNet-x1	85.26
DeepOC-SORT++	GT boxes	OSNet-x0.75	87.85

Table 4. Ablation study of DeepOC-SORT's different components on SoccerNet train set: (post) merging post processing, fine-tuned player detector (det), fine-tuned appearance model (app) and GSI + AFLink (++).

post	det	app	++	HOTA ↑
				57.3
			✓	59.0
		✓	✓	60.1
	✓	✓	✓	66.0
✓	✓	✓	✓	66.4

References

[1] 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. 1

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 1

[3] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 1

[4] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*, 2022. 1

[5] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1

[6] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv preprint arXiv:2302.11813*, 2023. 1

[7] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023. 2

[8] 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

[9] 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. *International journal of computer vision*, 129:548–578, 2021. 3

[10] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 3