Observation Centric and Central Distance Recovery for Athlete Tracking

Hsiang-Wei Huang[1], Cheng-Yen Yang[1], Samartha Ramkumar[1], Chung-I Huang[2],
Jenq-Neng Hwang[1], Pyong-Kun Kim[3], Kyoungoh Lee[3], Kwangju Kim[3]

[1] Department of Electrical and Computer Engineering, University of Washington, Seattle, United States
[2] National Center for High-Performance Computing, Hsinchu, Taiwan
[3] Electronics and Telecommunications Research Institute, Daejeon, Korea

{hwhuang, cycyang, sr74, hwang}@uw.edu,
{1203033}@narlabs.org.tw, {iros, longweek7, kwangju}@etri.re.kr

Abstract

Multi-Object Tracking on humans has improved rapidly with the development of object detection and re-identification algorithms. However, multi-actor tracking over humans with similar appearance and non-linear movement can still be very challenging even for the state-of-the-art tracking algorithm. Current motion-based tracking algorithms often use Kalman Filter to predict the motion of an object, however, its linear movement assumption can cause failure in tracking when the target is not moving linearly. And for multi-player tracking over the sports field, because the players on the same team are usually wearing the same color of jersey, making re-identification even harder both in the short term and long term in the tracking process. In this work, we proposed a motion-based tracking algorithm and three post-processing pipelines for three sports including basketball, football, and volleyball, we successfully handle the tracking of the non-linear movement of players on the sports fields. Experimental results achieved a HOTA of 73.968 on the testing set of ECCV DeeperAction Challenge SportsMOT Dataset and a HOTA of 49.97 on the McGill HPTDataset, showing the effectiveness of the proposed framework and its robustness in different sports including basketball, football, hockey, and volleyball.

1. Introduction

Tracking is a fundamental task in computer vision, aiming to associate objects and keep track of all the identities in video sequences. The recent development in deep learning pushes the performance of tracking with the aid of more accurate bounding boxes and feature extraction [13, 17, 20]. A great portion of multi-object tracking datasets focus on pedestrians in crowded street scenes (e.g., MOT16, MOT17, and MOT20) [10]. In these scenes, most of the pedestrians' movement is slow and linear and thus predictable; besides that, the appearance of each identity is easy to distinguish, making the re-identification in the tracking process more accessible. However, there is a lack of success in multi-object tracking algorithms that can successfully handle the sports scene where multiple players need to be tracked on the field. There are two main challenges posed: (1) The non-linear movement of sports players and (2) The similar appearances between players on the sports field scenes. For this purpose, we introduced an observation-centric and central distance recovery tracking algorithm that can handle the non-linear movement of players on the sports field and incorporate the appearance ReID post-processing to deal with the fragmented tracklets during tracking.

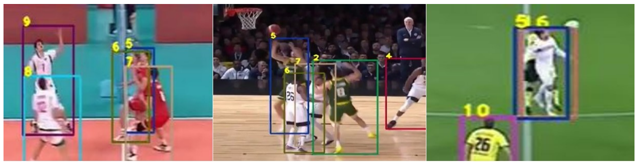

Figure 1. Occlusion can occur during the tracking process of sports players, which can cause ID switch and lower the tracking performance.

2. Related Works

2.1. Appearance-Based Object Tracking

With the fast development in the appearance feature extractor, some tracking algorithms incorporate target appearance during association [13, 17] and utilizes the appearance as a clue for identity recognition. However, most of these appearance cues-based algorithms often fall short in many cases, especially when targets are occluded, when the scene is very crowded, or the targets are sharing similar appearances.

2.2. Location and Motion-Based Object Tracking

Modern object-tracking algorithms usually follow the paradigm of tracking by detection. Several motion-based tracking algorithms adopt the Kalman filter [5] to formulate the moving trajectories of the target with the detections from the object detector. However, the linear assumptions of the Kalman filter can fall short when the target is not moving linearly. Observation-Centric SORT [3] includes moving direction similarity between detection and tracklets into the Hungarian Algorithm [7] cost and reforming the trajectories of the re-identify target to prevent error accumulations in the Kalman filter update process. These methods show effectiveness when dealing with the non-linear movements of objects, achieving state-of-the-art tracking performance on several multi-object tracking datasets.

2.3. Multi-People Tracking for Sports

Various works have been designed to serve the purpose of tracking team sport players on the field throughout the game. Such tracking not only can aid the automation of game statistic recording but allow sports analytics to gather high-level information from a video scene understanding perspective. Most works adapt tracking-by-detection method and incorporate an additional re-identification network to produce the embedding feature for each player.

Vats et al.[12] demonstrated a combination of team-identification and player-identification branch to assist the tracking process in hockey. Yang et al.[14] and Maglo et al. [9] shown that with the localization of the field and players, the tracking results will be much more robust in football. Sangüesa et al. [11] further introduces the idea of bridging the semantics of poses and actions to the embedding features to enhance the tracking in basketball.

3. Proposed Method

Several challenges need to be tackled in the tracking of sports players. First, is the non-linear movement of the players on the court. Given the high intensity of sports like volleyball, basketball, and football, players need to sprint, jump and change directions in a short time during the game, causing the movement to be unpredictable. Second, is the heavy occlusion problem during the sports game. The players will cluster together during the sports game when some particular situations happened like grabbing the rebound in basketball or blocking in volleyball. When this happened, it causes detection to be unreliable due to occlusion, and thus causes the tracking performance to drop if we do not take care of the recovery of the tracklets with lost detection carefully. Third, in the sports video clips, a player can go out and re-enter the camera field of view again after several seconds, when this happened, we need to re-identify the player with the correct identity. However, given the similar appearance between those team players within the same team, it is difficult to re-identify them correctly by using their appearance. In our work, we proposed several methods to deal with these three challenges.

3.1. Observation-Centric SORT

The non-linear movement is troublesome during the tracking process. To tackle this problem, we adopt observation-centric SORT (OCSORT)[3] as our main tracking algorithm and make reasonable modifications based on the task we try to tackle. Its observation-centric online smoothing strategy helps deal with the non-linear movement of the target. Once an object is being untracked and re-associate after a certain time, for example, the last observation before untracked is z_{t_1}, and the observation triggering re-association is z_{t_2}, we can conduct observation online centric smoothing by rebuilding a virtual trajectory $T_{virtual}$ following:

$$\hat{z} = T_{virtual}(z_{t_1}, z_{t_2}, \mathrm{t}) \tag{1}$$

where $t_1 < t < t_2$.

Observation-centric online smoothing can start from t_1 to backcheck the parameters of the standard Kalman filter [6] by alternating between its prediction and update stages when a lost target is associated again along this virtual trajectory. By backchecking the Kalman filter's parameters, observation-centric online smoothing can prevent the error accumulation caused by occlusion or lost detection. Because the players on the sports field can change moving direction in a short time, OCSORT introduced object momentum(observation-centric momentum) in the association cost of tracking by detection process, which also helps reduce the tracking error caused by sudden direction changes. By introducing a momentum term in the Hungarian matching stage, the tracking algorithm can have more tolerance for the non-linear movements of objects, improving the tracking robustness and reducing the noise. Given N existing tracklets and M unassociated detections, the associated cost C for the Hungarian assignment is calculated in pair-wise fashion:

$$C_{ij} = C_{ij}^{IoU} + \lambda C_{ij}^{m} = f_{IoU}(\hat{X}_i, Z_j) + \lambda f_m(\hat{X}_i, Z_j, V_i) \tag{2}$$

where $\hat{\mathrm{X}} \in \mathbb{R}^{N\times 7}$ are the sets of estimated object states and $\mathrm{Z} \in \mathbb{R}^{M\times 5}$ are detections. $\mathrm{V} \in \mathbb{R}^{N}$ calculates the moving directions of existing tracks with two previous observations of time difference Δt. The cost is a combination of f_{IoU} and f_m, where f_{IoU} stands for the negative pairwise Intersection over Union for every detection and the Kalman filter prediction bounding box of each tracklet. f_v calculates the moving direction consistency between the tracklet's moving directions and the direction formed by a track's last observation and the detection. λ is a weighting factor between the IoU term and the momentum term.

Lastly, after the association stage based on Kalman filter prediction is finished, OCSORT performs observation-centric recovery, trying to associate the last observation of an unassociated track to the detections based on their IoU. This strategy helps to reduce the generation of a new tracklet, which is usually abnormal given the fixed number of players on the sports field.

3.2. Central Distance Recovery

After the association step, there is still a chance that observation-centric recovery failed to successfully recover the identities of the targets and initiate new tracklets. This is mainly due to observation-centric recovery being based on using bounding box IoU as the Hungarian assignment cost, while the unassociated tracklets and the unmatched detections sometimes do not necessarily share an overlapping region given the fast-moving speed of sports players. To deal with this, we use the Euclidean distance between detections and the last observation of unassociated tracklets to conduct observation-centric recovery again, as this stage is based on bounding boxes' central distance, we call this process "central distance recovery".

3.3. Tracklets Post-Processing

The tracking results usually end up with more identities than the actual number of identities. This is mainly due to when a player re-enters the camera view, the tracking algorithm usually treats the player as a new identity and does not re-identify the player. To deal with the player re-entry problem, we incorporate different strategies in three sports scenes according to the total number of sports players, the size of the sports field, and several other sports characteristics.

Post-Processing in Basketball. The total number of players in a basketball game is 10, we use this as the main constraint to conduct the post-processing. Due to the quick camera horizontal movement and the nature of basketball, there are a large amount of player's camera re-entries compared to other sports, it is necessary to incorporate person re-identification in the post-processing stage. After we get the preliminary tracking result, we initiate the first 10 tracklets we get as the ten main players on the court. We keep updating the tracklets appearance feature using exponential moving averages. Whenever a player leaves the camera view, we put the player's tracklet into a candidate queue. And whenever a new tracklet (player) enters the camera view, we associate the player with one of the tracklets in the candidate queue that shares the highest cosine similarity between both.

Post-Processing in Football. The search space of candidates in football is much bigger compared to other sports because of the bigger number of players in a football game and the lower ratio of players inside the camera view to the total number of players. Thus it is necessary to incorporate the position as a constraint during the ReID process. In football players' tracking, due to the uncertainty in the number of identities in a video clip, we decided not to use the number of total players as a constraint, instead, we only try to associate fragment tracklets together based on their appearance similarity and entry/exit position. There is a total of three rounds of association based on appearance similarity and location between tracklets. The prerequisites for two tracklets to be associated together are first based on their disappear and reappear location. The threshold for the location distance for two tracklets to be associated is determined by the number of frames between a tracklet disappearing and reappearing in the camera view, which means that if a tracklet disappears and reappears in the camera view in a short amount of time, the distance threshold for their location will be small given the moving distance should not be far in such a short time, and vice versa for a longer disappearing time. After passing the threshold of location distance, we then try to compare the cosine similarity based on the two tracklets' appearance. We calculate two tracklets' average frame-based embedding features distance as their final distance. If the final distance is smaller than the embedding threshold, we consider these two tracklets the same identity. We use three rounds of association based on a different matching threshold of appearance similarity.

Post-Processing in Volleyball. Compared to basketball and football, volleyball player usually stays in the camera view throughout the entire video clip, and even they disappear, they reappear in a very close location. Due to the above reason, we do not incorporate appearance in the post-processing of volleyball. We use a similar strategy to basketball post-processing, we limited the number of tracklets to 12 players, and try to re-associate disappearing players to candidates only based on the distance of their disappear and reappear location.

After the re-identification post-processing part is done, we use standard linear interpolation as our last step to produce the final tracking results.

4. Experiments and Results

4.1. Datasets

4.1.1 SportsMOT Dataset

For the algorithm evaluation, We use the training sets from SportsMOT for detector and ReID model training. The training set contains 45 video clips from 3 categories (i.e., basketball, football, and volleyball), which are collected

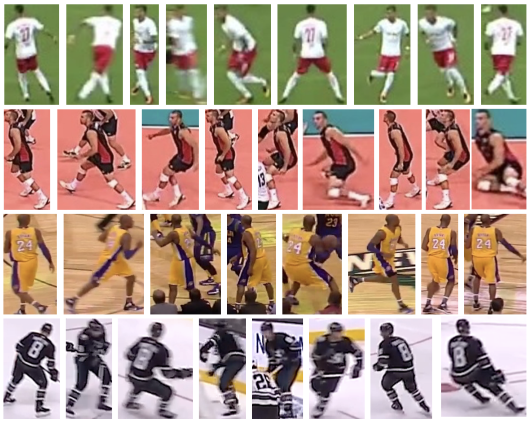

Figure 2. Some samples for the SportsMOT Dataset and MGill Hockey Player Tracking Dataset. We can observe some blurriness even with a very high FPS, under the condition of a fast-moving camera and players.

Sport Type	# of tracks	# of frames	Track Len	Density
Basketball	10	845.4	767.9	9.1
Football	22	673.9	422.1	12.8
Volleyball	12	360.4	335.9	11.2

Table 1. Summary of the SportsMOT dataset split by the type of sport. The number of tracks, number of frames, track length, and track density are average numbers across all videos of the sport.

from Olympic Games, NCAA Championship, and NBA games on YouTube. Only the search results with 720P resolution, 25 FPS, and official recordings are downloaded. Note that only players that are actively on the game field are tracked, excluding the referees, coaches, and benched players. No substitutions or any kind of player exchanges are in the sequences.

4.1.2 McGill Hockey Player Tracking Dataset

Another dataset used in our experiment is the McGill Hockey Player Tracking Dataset[15], the dataset consists of 25 National Hockey League gameplay video clips. Each clip contains one shot of the gameplay from the overhead camera position without any cut or camera switch. For the National Hockey League broadcast video, there are two popular video frame rates that are available on the market, 30 fps and 60 fps. In this dataset, half the video clips have a frame rate of 30 fps, and the other half have a frame rate of 60 fps. Different from the SportsMOT dataset, the hockey player and the camera movement are faster and more unpredictable. Also, the players within the same team are sharing more similar appearances, thus making appearance-based re-identification even harder.

4.2. Implementation Details

4.2.1 Detector

We use YOLOX[4] as our detector due to its high accuracy and fast inference speed. For the pretrained weight, we use the COCO pretrained YOLOX-X model provided by the official GitHub repositories of YOLOX. We train the model with Sportsmot training set for 80 epochs, following the YOLOX-X default training process of ByteTrack's [8] official GitHub repositories. The training duration takes around 8 hours on 4 Nvidia Tesla V100 GPUs.

4.2.2 Tracking Algorithm

Observation-Centric SORT. We keep the original configuration of OCSORT, using 0.1 detection confidence threshold, 0.3 IoU threshold, 0.7 track threshold, and a maximum tracklet age of 30 frames for all of the sports.

Central Distance Recovery The central distance recovery threshold is different based on the sports type. We set basketball's distance threshold as 200, football's distance threshold as 80, volleyball's distance threshold as 80 and hockey's distance threshold as 50. The choice of threshold is based on the evaluation performed on the testing set.

4.2.3 Player Re-Identification

For player re-identification, we followed the omni-scale feature learning proposed in OSNet [19], using the unified aggregation gate to fuse the features from different scales. We adapt the weights trained on Market1501 [18] and then fine-tuned the model on SportsMOT Re-ID dataset. SportsMOT Re-ID dataset is constructed based on the original SportsMOT dataset where we cropped out each player according to its ground-truth annotation of the bounding boxes. The first 25 frames of each track are chosen as the query set while the remaining are selected as the gallery set.

To allow our player re-identification model to have more generalization ability, we took all three sports into training in this step. The backbone of the model is ResNet-50 and is trained for 10 additional epochs with Adam optimizer and a 0.0003 learning rate.

4.2.4 Post-Processing

Basketball Setting. In the post-processing of basketball, we limited the number of tracklets to 10, just like the number of players on the court, unless more than 10 detections appear at the same time, we do not initiate new tracklets. The re-identification of players is based on the cosine similarity of their embedding features.

Football Setting. In football, considering the ratio of players inside and outside the camera, we use the cosine similarity and also players' position to conduct re-identification for

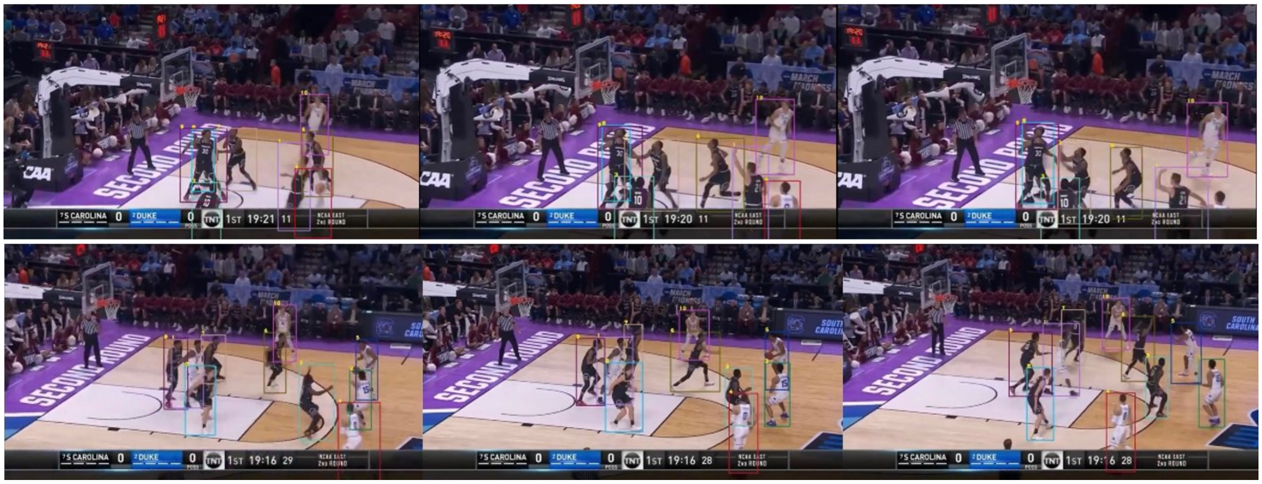

Figure 3. The red bounding box with tracking ID 4 left the camera view in the first row of demo image, and the post-processing retrieved its tracking ID when he re-entered the camera view in the second row of demo image after four seconds.

Ranking	Team	HOTA↑	MOTA↑	IDF1↑	IDs↓	Frag↓
1	Team BOE_AIoT_CTO	**76.254**	89.316	**84.453**	**2567**	6104
2	Team IPIU	74.899	95.590	78.342	4853	4536
3	Ours	73.968	94.832	78.271	2754	**3592**
4	Team AI PRIME	73.225	**96.018**	73.963	3405	4123
5	Team MiaoMiao	71.489	83.632	73.862	11271	11949
6	Team xiaochangcheng	70.989	94.446	71.883	3085	3849

Table 2. The SportsMOT competition final leaderboard. Our final performance ranked 3rd place on HOTA among 220 valid submissions. The best performance on each evaluation metric is in **bold face**. With our observation-centric and central distance recovery method, we are able to minimize the number of both ID switches and fragments.

camera re-entry players. There are three rounds of the association stage. The association between two tracklets needs to pass through a threshold of tracklet position distance before they have a chance to be associated. The distance threshold is based on their disappear and reappear time gap, for those tracklets that have a time gap of fewer than 100 frames, the distance threshold is 100. For the tracklets that share a time gap between 100 to 500 frames, the distance threshold is 250, and for those tracklets that share a time gap bigger than 500 frames, the distance threshold is set to 400. In three rounds of association, we try to associate as many as tracklets we can in a greedy style. For the first round of association, if the cosine distance of two tracklets is smaller than 0.1, we treat them as the same identity and conduct association. For the second round, the cosine distance threshold is 0.2, and for the third round is 0.4.

Volleyball Setting. Due to the relatively low number of player camera re-entry cases in volleyball compared to basketball and football, the search space for re-entry players is small. So the post-processing of volleyball is simply based on their disappear and reappear position. For the re-entry player, we select the player in the re-associate candidate who has the closest distance between the disappearing position of candidates and reappears position of the new player for re-identification if their distance is lower than a threshold of 400.

Hockey Setting. Due to the high appearance similarity of hockey player in the same team, we do not use appearance re-identification as post-processing. Also, in ice hockey, players are substituted ”on the fly,” meaning a substitution can occur even in the middle of play, which makes the trick of limiting the total number of identities on the court useless because the identities on the court can change anytime. Thus, we do not incorporate any re-entry handling post-processing for now.

Method	HOTA	MOTA	IDF1
ByteTrack	59.76	93.79	64.85
OCSORT	67.11	92.95	65.83
OCSORT+CDR	71.19	93.81	74.39
OCSORT+CDR+Post Processing	**73.97**	**94.83**	**78.27**

Table 3. The performance of different methods on the SportsMOT test set. CDR stands for central distance recovery.

4.3. Evaluation

The 2022 ECCV DeeperAction Challenge - SportsMOT Track on Multi-actor Tracking competition is ranked according to the HOTA[8] performance. In contrast to MOTA[1], HOTA maintains a balance between the accuracy of object detection and association. The original OCSORT has a performance of 67.107 in HOTA, after the incorporation of central-distance recovery, the HOTA improves to 71.764. After the ReID post-processing stage, we achieve 73.968 in HOTA, 63.460 in AssA, 86.316 in DetA, 94.832 in MOTA, 78.271 in IDF1, 2754 in IDS, and 3592 in Frag, showing the effectiveness of our method.

To demonstrate our algorithm's generalization ability on different team sports, we evaluate our algorithm on the McGill Hockey Player Tracking Dataset [15]. In contrast to the previous three sports, tracking the hockey players is more challenging due to most of the players' appearance (hair color or skin tone) being occluded with hockey gear with identical color and can not be re-identified based on their appearance features. Thus, we do not include any extra stage of appearance-based post-processing and just evaluate the performance of OCSORT and central distance recovery. We evaluate the performance of our algorithm and compare it to several other popular motion-based tracking algorithms[2, 16]. The experiment results showing our algorithm outperform others in HOTA, IDF1, and MOTA, demonstrating the robustness in different types of sports players tracking.

Method	HOTA	MOTA	IDF1
ByteTrack	47.61	73.78	58.02
OCSORT	48.74	74.91	59.13
OCSORT+CDR	**49.97**	**76.31**	**61.32**

Table 4. The performance of different methods on the McGill Hockey Player Tracking Dataset test set.

Method	HOTA	MOTA	IDF1
ByteTrack	76.25	-	73.45
OCSORT	80.64	-	73.51
OCSORT+CDR	**82.81**	-	**75.58**

Table 5. The performance of different methods using ground truth detection on the McGill Hockey Player Tracking Dataset test set.

5. Conclusions

In this paper, we modify the motion-based observation-centric SORT with an extra central distance recovery stage, improving the performance without adding too much computational cost and also keeping the algorithm online, successfully tackle down the challenges of non-linear movement during tracking. We also propose three different post-processing for each sport according to the sports characteristics. Our final performance achieves 73.968 in HOTA on the ECCV DeeperAction Challenge - SportsMOT dataset, outperforming most of the teams joined in the competition.

6. Acknowledgement

This research was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (22ZD1120, Development of ICT Convergence Technology for Daegu-GyeongBuk regional Industry). We also want to acknowledge and thank National Center for High-Performance Computing from Taiwan for providing the computing resources.

References

[1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics, 2008. EURASIP Journal on Image and Video Processing, 2008:1–10.

[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, 2016.

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

[4] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021, 2021. arXiv preprint arXiv:2107.08430.

[5] R. E. Kalman. A new approach to linear filtering and prediction problems, 1960. J. Fluids Eng., 82(1):35–45.

[6] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[7] Harold W Kuhn. The hungarian method for the assignment problem, 1955. Naval research logistics quarterly, 2(1-2):83–97.

[8] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian ´ Leibe. Hota: A higher order metric for evaluating multi-object tracking, 2021. International journal of computer vision, 129(2):548–578.

[9] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Efficient tracking of team sport players with few game-specific annotations. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3460–3470, 2022.

[10] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking, 2016. arXiv preprint arXiv:1603.00831.

[11] Adrià Arbués Sangüesa, Coloma Ballester, and Gloria Haro. Single-camera basketball tracker through pose and semantic feature fusion. *CoRR*, abs/1906.02042, 2019.

[12] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A Clausi, and John S. Zelek. Player tracking and identification in ice hockey. *ArXiv*, abs/2110.03090, 2021.

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

[14] Yukun Yang, Ruiheng Zhang, Wanneng Wu, Yu Peng, and Min Xu. Multi-camera sports players 3d localization with identification reasoning. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4497–4504, 2021.

[15] Kua Chen Yingnan Zhao, Zihui Li. A method for tracking hockey players by exploiting multiple detections and omni-scale appearance features. *Project Report*, 2020.

[16] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Byte-track: Multi-object tracking by associating every detection box, 2021. arXiv preprint arXiv:2110.06864.

[17] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, , and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking, 2021. International Journal of Computer Vision, 129(11):3069–3087.

[18] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jing-dong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[19] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification, 2019. Proceedings of the IEEE/CVF International Conference on Computer Vision.

[20] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ArXiv*, abs/2004.01177, 2020.