Iterative Scale-Up ExpansionIoU and Deep Features Association for Multi-Object Tracking in Sports

Hsiang-Wei Huang[1], Cheng-Yen Yang[1], Jiacheng Sun[1], Pyong-Kun Kim[2]
Kwang-Ju Kim[2], Kyoungoh Lee[2], Chung-I Huang[3], Jenq-Neng Hwang[1]
[1]Information Processing Lab, University of Washington
[2]Electronics and Telecommunications Research Institute
[3]National Center for High-Performance Computing
{hwhuang,cycyang,sjc042,hwang}@uw.edu
{iros,kwangju,longweek7}@etri.re.kr
1203033@narlabs.org.tw

Abstract

Deep learning-based object detectors have driven notable progress in multi-object tracking algorithms. Yet, current tracking methods mainly focus on simple, regular motion patterns in pedestrians or vehicles. This leaves a gap in tracking algorithms for targets with nonlinear, irregular motion, like athletes. Additionally, relying on the Kalman filter in recent tracking algorithms falls short when object motion defies its linear assumption. To overcome these issues, we propose a novel online and robust multi-object tracking approach named deep ExpansionIoU (Deep-EIoU), which focuses on multi-object tracking for sports scenarios. Unlike conventional methods, we abandon the use of the Kalman filter and leverage the iterative scale-up ExpansionIoU and deep features for robust tracking in sports scenarios. This approach achieves superior tracking performance without adopting a more robust detector, all while keeping the tracking process in an online fashion. Our proposed method demonstrates remarkable effectiveness in tracking irregular motion objects, achieving a score of 77.2% HOTA on the SportsMOT dataset and 85.4% HOTA on the SoccerNet-Tracking dataset. It outperforms all previous state-of-the-art trackers on various large-scale multi-object tracking benchmarks, covering various kinds of sports scenarios. The code and models are available at https://github.com/hsiangwei0903/Deep-EIoU.

1. Introduction

Multi-Object Tracking (MOT) is a fundamental computer vision task that aims to track multiple objects in a video and localize them in each frame. Most recent tracking algorithms [33, 1, 28, 4], which mainly focus on pedestrians or vehicle tracking, have achieved tremendous progress on public benchmarks [19, 8, 11]. However, these state-of-the-art algorithms fail to perform well on datasets with higher difficulties, especially those datasets with sports scenarios [7, 6, 36]. Given the growing demand for sports analytic for applications like automatic tactical analysis and athletes' movement statistics including running distance, and moving speed, the field of multi-object tracking for sports requires more attention.

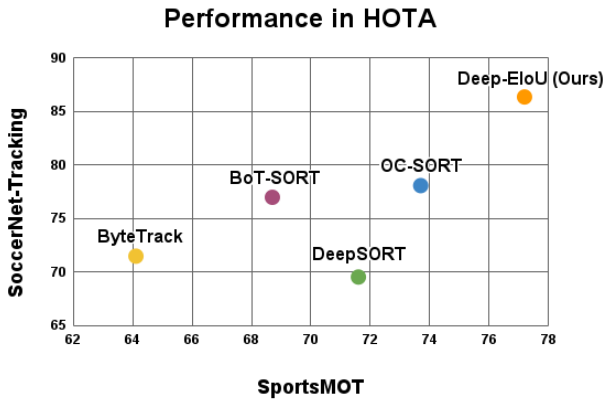

Figure 1. HOTA comparison of different trackers on the test sets of SoccerNet-Tracking and SportsMOT dataset. Deep-EIoU achieves 77.2% HOTA on the SportsMOT test set and 85.4% HOTA on the SoccerNet-Tracking test set. These results surpass the performance of all previous trackers on these large-scale multi-object tracking benchmarks. More comparisons between different trackers can be found in table 2 and table 3

Different from multi-object tracking for pedestrians or vehicles, MOT in sports scenarios poses higher difficulties

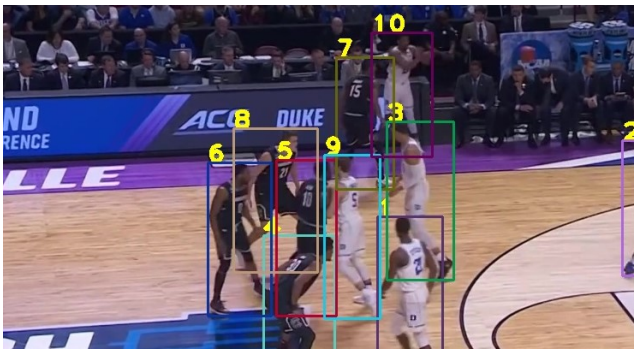

Figure 2. An example of the occlusion problem encountered during multi-athlete tracking. Occlusion can significantly hinder detection and tracking performance, and the occlusion issue in athlete tracking is particularly severe when compared to pedestrian tracking due to the high intensity of sports characteristics.

due to several reasons, including severe occlusion caused by the high intensity in sports scenes as illustrated in Figure 2, similar appearance between players in the same team due to the same color jersey like examples in Figure 3, and also unpredictable motion due to some sport movement like a crossover in basketball, sliding tackle in football or spike in volleyball. Due to the above reasons, the previous trackers, which utilize appearance-motion fusion [34, 28] or simply motion-based [33, 5, 4] methods struggle to conduct robust tracking on several major MOT benchmarks in sports scenarios [6, 7].

To address these issues, in this paper, we propose a novel and robust online multi-object tracking algorithm specifically designed for objects with irregular and unpredictable motion. Our experimental results demonstrate that our algorithm effectively handles the irregular and unpredictable motion of athletes during the tracking process. It outperforms all tracking algorithms on two large-scale public benchmarks [7] without introducing extra computational cost while maintaining the algorithm online. Therefore, in this paper, we assert three main contributions:

- We present a novel association method to specifically address the challenges in sports tracking, named ExpansionIoU, which is a simple yet effective method for tracking objects with irregular movement and similar appearances.
- Our proposed iterative scale-up ExpansionIoU further leverages with deep features association for robust multi-object tracking for sports scenarios.
- The proposed method achieves **77.2** HOTA on the SportsMOT [7] dataset, and **85.4** HOTA on the SoccerNet-Tracking dataset [6], outperforming all the other previous tracking algorithms by a large margin.

Figure 3. Example of similar appearances between the players from the SportsMOT dataset, which can cause confusion towards the tracker and decrease the tracking accuracy. Each column represents two different players with similar appearance.

2. Related Work

2.1. Multi-Object Tracking using Kalman Filter

Most of the existing tracking algorithms [33, 4, 5, 28, 35, 30, 14, 12, 13, 29] incorporate Kalman filter [15] as a method for object motion modeling. Kalman filter can formulate object motion as a linear dynamic system and can be used to predict its next frame location according to the object's motion from the previous frames. Kalman filter has shown effectiveness in multi-object tracking across several public benchmarks [19, 8, 23]. However, due to the Kalman filter's linear motion and Gaussian noise assumption, the Kalman filter might fail to track an object with non-linear motion. Due to this reason, OC-SORT [5] proposes several methods including observation-centric re-update to modify the Kalman filter's parameters during the tracking process and prevent error accumulations when an object is not tracked. The performance has shown effectiveness for tracking objects with irregular motion on several public datasets [23, 7].

2.2. Location-based Multi-Object Tracking

Tracking can also be conducted based on the position information, given a high frame rate input video sequence, the object's position shift between frames is relatively small due to the high frame rates, thus making the position information a reliable clue for association between frames. Several methods [22, 14] utilizes the bounding boxes' distance as the cost for bounding box association, while some recent work [31] utilize different IoU calculation methods including GIoU [20], DIoU [38], and BIoU [31], to conduct bounding box association between frames, which also demonstrate effectiveness in multi-object tracking.

2.3. Appearance-based Multi-Object Tracking

With the recent development and improvement of object ReID model [39] and training tricks [17], many tracking algorithms incorporate ReID into the association process. Some methods use the joint detection and embedding architecture [35, 27] to produce detection and object embedding at the same time to achieve real-time tracking.

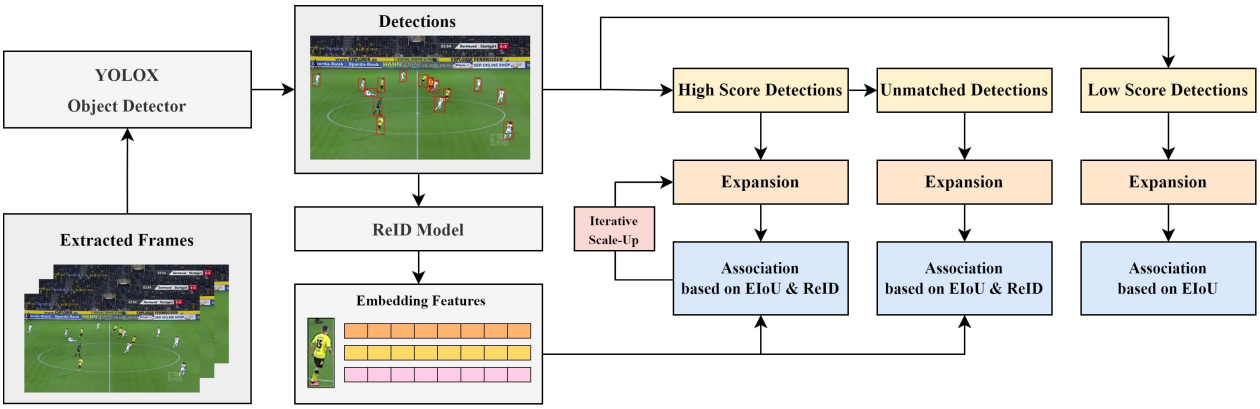

Figure 4. The proposed iterative scale-up ExpansionIoU tracking pipeline. The pseudo code of the proposed pipeline can be found in supplementary material.

While the other methods [28, 1] apply other stand-alone ReID model to extract detection's embedding features for association. The appearance-based tracking methods improve the tracking robustness with an extra appearance clue, while sometimes the appearance can be unreliable due to several reasons including occlusions, similar appearance among tracked objects, appearance variation caused by the object's rotation, or the lighting condition.

2.4. Multi-Object Tracking in Sports

Numerous studies have been conducted to monitor players' movements in team sports during games. This monitoring serves not only to automate the recording of game statistics but also enables sports analysts to obtain comprehensive information from a video scene understanding perspective. Different from MOT of pedestrian [19], MOT in sports scenarios is much more challenging due to several reasons including targets' faster and irregular motions, similar appearance among players in the same team, and more severe occlusion problem due to the sport's intense characteristic. The majority of recent methods for MOT for sports utilize the tracking-by-detection paradigm and integrate a re-identification network to generate an embedding feature for association.

Vats et al. [25] combine team classification and player identification approaches to improve the tracking performance in hockey. Similarly, Yang et al. [32] and Maglo et al. [18] demonstrate that by localizing the field and players, the tracking results in football can be more accurate. Additionally, Sangüesa et al. [21] utilize the human pose information and actions as the embedding features to enhance basketball player tracking. While Huang et al. [14] combine OC-SORT [5] and appearance-based post-processing to conduct tracking on multiple sports scenarios including basketball, volleyball, and football [7].

3. Proposed Methods

Our proposed method follows the classic tracking-by-detection paradigm, which also enables online tracking without using future information. We first apply the object detector YOLOX on each input frame, and then we conduct association based on several clues including the similarity between extracted appearance features and the Expansion-IoU between the tracklets and detections. After the association cost is obtained, the Hungarian algorithm is conducted to get the best matching between tracklets and detections.

3.1. Appearance-based Association

The appearance similarity is a strong clue for object association between frames, the similarity can be calculated by the cosine similarity between the appearance features, and it can also be used to filter out some impossible associations. The cost for appearance association $Cost_A$ can be directly obtained from the cosine similarity with the following formula:

$$Cost_A = 1 - \text{Cosine Similarity} = 1 - \frac{a \cdot b}{\|a\|\|b\|} \quad (1)$$

Here, a and b are the tracklet's appearance feature and the detection's appearance feature, respectively. A higher cosine similarity denotes a higher similarity in appearance, while a lower cosine similarity means the tracklet's appearance and the detection's appearance are different.

3.2. Association with ExpansionIoU

Insipired by previous work [31], which utilizes expanding bounding boxes for association, to deal with the fast and irregular movement of sports player, we proposed ExpansionIoU (EIoU), a robust association method for tracking under large and nonlinear motion. Different from the previous work [31], we found out that expanding the bounding box even more during association can lead to a significantly

better performance in athlete tracking. Traditional IoU has been a cornerstone in location-based tracking method, but it often lacks the flexibility to account for object's large movement, when tracklet and detection bounding boxes share small or no IoU between adjacent frames. EIoU addresses this limitation by modifying the dimensions of bounding boxes, expanding their width and height and considers a wider range of object relationships, thus recover the association for those objects with large movement in sports scenarios. The expansion of bounding box is controlled by expansion scale E, given an original bounding box with height h and width w, we can calculate the expansion length $h^\star$ and $w^\star$ following:

$$\begin{aligned} h^\star &= (2E+1)h \\ w^\star &= (2E+1)w \end{aligned} \tag{2}$$

The original bounding box is expand based on the expansion length. Denote the original bounding box top-left and bottom-right coordinate as (t,l),(b,r), we can derive the expanded bounding box's coordinate as $(t-\frac{h^\star}{2}, l-\frac{w^\star}{2})$ and $(b+\frac{h^\star}{2}, r+\frac{w^\star}{2})$.

The expanded bounding box is further used for IoU calculation between tracklets and detections pairs, note that the expansion is applied both on tracklets' last frame detections and the new coming detections from detector, the calculated EIoU is used for Hungarian association between adjacent frames. The operation of expanding the bounding box does not change several important objects' information like the bounding box center, aspect ratio, or appearance features. By simply expanding the search space, we can associate those tracklets and detections with small or no IoU, which is considered a common situation when the target's movement is fast, especially in sports games.

3.3. Confidence Score Aware Matching

Following ByteTrack [33], we give the high confidence score detections higher weighting during the matching process. The high score detections usually imply less occlusion, hence a higher chance to preserve more reliable appearance features. Due to this reason, the first stage matching with high score detections is based on the association cost of both appearance and ExpansionIoU, denoted as C_{stage1}. The first stage of matching is built upon several rounds of iterative associations with a gradually scale-up expansion scale, addressed in Section 3.4. In the second round of matching with low score detections, only ExpansionIoU is used, the cost is denoted as C_{stage2}.

In our first matching stage, we abandon the IoU-ReID weighted cost method used in several previous works [34, 28], where the cost is a weighted sum of the appearance cost C_A and IoU cost C_{IoU}:

$$C = \lambda C_A + (1-\lambda) C_{IoU} \tag{3}$$

Instead, we adopt strategy similar to that of BoT-SORT [1] for appearance-based association. More specifically, we first filter out some impossible associations by setting cost thresholds for both appearance and ExpansionIoU (EIoU). The adjusted appearance cost $C_{\hat{A}}$ is set to 1 if either cost is bigger than its corresponding threshold, otherwise $C_{\hat{A}}$ is set as half of its appearance cost C_A. Finally, the first stage's final association cost C_{stage1} is set as the minimum of the appearance cost $C_{\hat{A}}$ and EIoU cost C_{EIoU}. With τ_A and τ_{EIoU} denotes the threshold for the cost filter, we can write the appearance cost $C_{\hat{A}}$ as:

$$C_{\hat{A}} = \begin{cases} 1, & \text{if } C_A > \tau_A \text{ or } C_{EIoU} > \tau_{EIoU} \\ 0.5 C_A, & \text{otherwise} \end{cases} \tag{4}$$

The final cost in the first stage of matching C_{stage1} will be the minimum between adjusted appearance cost $C_{\hat{A}}$ and EIoU cost C_{EIoU}.

$$C_{stage1} = \min(C_{\hat{A}}, C_{EIoU}) \tag{5}$$

While the association cost in the second matching stage C_{stage2} will be only using the EIoU cost C_{EIoU}.

3.4. Iterative Scale-Up ExpansionIoU

As illustrated by the previous work using expansion bounding box for association [31], the amount of the bounding box expansion is a crucial and sensitive hyperparameter in the tracking process and the performance of the tracker can be largely affected by the choice of the hyperparameter. In the real-world scenario, several factors might limit us from tuning the expansion scale and improving the tracking performance, including 1) the online tracking requirements. One common requirement for an athlete tracking system is the system needs to operate in an online matter, tuning the expansion scale with experiments and tweaking the performance is not possible in such cases. 2) No access to the testing data. For real-world scenarios, the testing data's ground truth is often not available, which makes finding the perfect expansion scale for association impossible. Due to the above reasons, we proposed a novel iterative scale-up ExpansionIoU association stage for robust tracking, the experiment results show that without any parameter tuning, our algorithms can always maintain SOTA performance on public benchmark. Instead of doing hyperparameter tuning for the best expansion scale E, we choose to iteratively conduct EIoU association based on a gradually increasing E_t during the tracking process. In each scale-up iteration, the expansion scale of the current iteration E_t can be derived from the following formula:

$$E_t = E_{initial} + \lambda t, \tag{6}$$

where $E_{initial}$ is the initial expansion scale, λ denotes the step size for the iterative scale-up process, t stands for the iteration count, which starts from 0. By using this approach, we can first perform association to those trajectory and detection pairs with higher ExpansionIoU, and gradually search for those pairs with lower overlapping area, which enhances the robustness of our association process. Note that the iterative scale-up process is only applied for high score detections association, once the iteration count reaches the total number of iteration t_{total}, the association for high score detections stops and the tracker moves on to the low score detections association stage.

4. Experiments and Results

4.1. Dataset

We evaluate our tracking algorithm on two large-scale multi-sports player tracking datasets, i.e., SportsMOT [7] and SoccerNet-Tracking [6].

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 sports.

SportsMOT consists of 240 video sequences with over 150K frames and over 1.6M bounding boxes collected from 3 different sports, including basketball, football, and volleyball. Different from the MOT dataset [19, 8], SportsMOT possesses higher difficulties including: 1) targets' fast and irregular motions, 2) larger camera movements, and 3) similar appearance among players in the same team.
SoccerNet-Tracking is a large-scale dataset for multiple object tracking composed of 201 soccer game sequences. Each sequence is 30 seconds long. The dataset consists of 225,375 frames, 3,645,661 annotated bounding boxes, and 5,009 trajectories. Unlike SportsMOT, which only focuses on the tracking of sports players on the court, the tracking targets of SoccerNet contains multiple object classes including normal players, goalkeepers, referees, and soccer ball.

4.2. Detector

We choose YOLOX [10] as our object detector to achieve real-time and high accuracy detection performance. Several existing trackers [33, 5, 1, 31] also incorporate YOLOX as detector, this also leads to a more fair comparison between these trackers with ours. We use the COCO pretrained YOLOX-X model provided by the official GitHub repositories of YOLOX [10] and further fine-tune the model with SportsMOT training and validation set for 80 epochs, the input image size is 1440 $\times$ 800, with data augmentation including Mosaic and Mixup. We use SGD optimizer with weight decay of 5 $\times$ 10$-$4 and momentum of 0.9. The initial learning rate is 10$-$3 with 1 epoch warmup and cosine annealing schedule, which follows the same training procedure of ByteTrack's [33]. As for the SoccerNet-Tracking dataset, since oracle detections are provided in the dataset, to make a fair comparison and focus on tracking, we directly use the oracle detections provided by the dataset for the evaluation of all trackers.

4.3. ReID Model

For player re-identification (ReID), we use the omni-scale feature learning proposed in OSNet [39]. The unified aggregation gate fuses the features from different scales and enhances the ability of human ReID.
SportsMOT The ReID training data for experiments on SportsMOT dataset is constructed based on the original SportsMOT dataset where we crop out each player according to its ground truth annotation of the bounding boxes. The sampled dataset includes 31,279 training images, 133 query images, and 1,025 gallery images.
SoccerNet-Tracking We sample the ReID training data from the SoccerNet-Tracking training set, we randomly select 100 ground truth bounding boxes for each player from randomly sampled videos, with 65 used as training images, 10 used as query images, and 25 used as gallery images. The sampled ReID data contains 7,085 training images, 1,090 query images, and 2,725 gallery images, with a total of 109 randomly selected identities.
Training Details We use the pre-trained model from the Market-1501 dataset [37] and further fine-tune the model based on each of the above mentioned sampled sports ReID datasets, resulting in two ReID models for these two datasets. Each model is trained for 60 epochs, using Adam optimizer with cross entropy loss and the initial learning rate is 3 $\times$ 10$-$4. All the experiments are conducted on single Nvidia RTX 4080 GPU.

4.4. Tracking Settings

The threshold for detection to be treated as high score detection is 0.6, while detections with confidence score between 0.6 and 0.1 will be treated as low score detections, the rest detections with confidence score lower than 0.1 will be filtered. The cost filter threshold τ_A and τ_{EIoU} are set to 0.25 and 0.5, respectively. We also remove the constraint of aspect ratio in the detection bounding box, since sports scenarios might have the condition when a player is lying on the ground, which is different from the MOT datasets where most of the pedestrians are standing and walking. For the high score detections association, the initial value of expansion scale $E_{initial}$ is set to 0.7 with a step size λ of 0.1, and

the total number of iteration t_{total} is 2. The expansion scale E for low score detections association is 0.7, while for unmatched detections is 0.5. The max frames for keeping lost tracks is 60. After tracking is finished, linear interpolation is applied to boost the final tracking performance.

4.5. Evaluation Metrics

MOTA [2] is often used as an evaluation metric for multi-object tracking task, however, MOTA mainly focuses on the detection performance instead of association accuracy. Recently, in order to balance between the detection and association performance, more and more public benchmarks start to use HOTA [16] as the main evaluation metric. For evaluation on the SportsMOT dataset, we adopt HOTA, MOTA, IDF1, and other associated metrics [3] for comparison. While for SoccerNet, we adopt HOTA metrics, with associated DetA, and AssA metrics, since only these metrics are provided by the evaluation server.

4.6. Performance

We compare our tracking algorithm with previous existing trackers on two large-scale multi-object tracking datasets in sports scenarios, the SportsMOT and SoccerNet-Tracking datasets. All the experiments are run on one Nvidia RTX 4080 GPU, and the tracking results are evaluated on the datasets' official evaluation server.

SportsMOT As shown in table 2, the performance of our proposed Deep-EIoU achieves **77.2** in HOTA, **79.8** in IDF1, **67.7** in AssA. The performance of our method achieves state-of-the-art results and outperforms all the other previous trackers while also keeping the tracking process in an online fashion, showing the effectiveness of our algorithm in multi-object tracking in sports scenarios.

SoccerNet To focus on the tracking performance and make a fair comparison, all the evaluated methods are using oracle detections provided by the SoccerNet-Tracking dataset [6]. The performance of our proposed method is reported in table 3. Our method achieves **85.443** in HOTA, **73.567** in AssA, **99.236** in DetA, which outperforms several state-of-the-art online tracking algorithms by a large margin. The performance of DeepSORT and ByteTrack are reported from the original SoccerNet-Tracking paper [6]. The competitive performance of Deep-EIoU in various large-scale sports player tracking datasets demonstrates the effectiveness of our algorithm in multi-object tracking in sports.

4.7. Ablation Studies on Deep-EIoU

In our experiments, Deep-EIoU is evaluated with different settings on the SportsMOT test set, including whether to incorporate appearance (ReID) during tracking, using iterative scale-up bounding box expansion, and using linear interpolation as post-processing. As shown in Table 4, after incorporating ReID model based on appearance association, the HOTA of Deep-EIoU is boosted by 3.8, showing that although sharing similar appearance between athletes, it is still important to use appearance as a clue for tracking in sport scenarios. With the iterative scale-up process (ISU), the gradually scale-up bounding box can first establish association with those tracklets and detections with higher EIoU, thus also increase the tracking performance, note that the iterative scale-up process is incorporate with a larger tracking buffer, unlike the default setting of 30 for pedestrian tracking, we use 60 due to the stronger occlusion characteristics of the sports scenarios. And finally, following most of the online tracking algorithm [33, 5], we also include linear interpolation (LI) as a strategy to boost the final tracking performance.

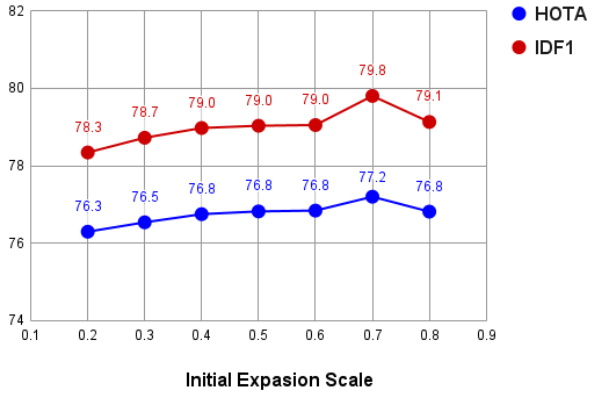

Figure 5. Performance comparison of Deep-EIoU under different initial expansion scales on the SportsMOT test set.

4.8. Robustness to initial expansion scale

To prove the effectiveness and robustness of our approach, we conduct experiments based on different initial expansion scales in the iterative scale-up process. We change the initial expansion scale from 0.2 to 0.8. The experiment results in Figure 5 show that we can still achieve SOTA performance with different initial expansion scales because the iterative scale-up process can enhance the robustness and does not require any parameter tuning to achieve SOTA performance. This proves our method's effectiveness in the real-world scenario, when ground truth is often not available and the tracking parameter can not be tuned.

4.9. ExpansionIoU on Kalman filter-based tracker

To test the effect of ExpansionIoU on the Kalman filter-based tracker, we also implement several versions of our method by directly incorporating the Kalman filter and ExpansionIoU. In our implementation, the Kalman filter's pre-

Method	Training Setup	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑	LocA↑	IDs↓	Frag↓
FairMOT [34]	Train	49.3	53.5	34.7	86.4	70.2	83.9	9928	21673
QDTrack [9]	Train	60.4	62.3	47.2	90.1	77.5	88.0	6377	11850
CenterTrack [40]	Train	62.7	60.0	48.0	90.8	82.1	90.8	10481	5750
TransTrack [24]	Train	68.9	71.5	57.5	92.6	82.7	91.0	4992	9994
BoT-SORT [1]	Train	68.7	70.0	55.9	94.5	84.4	90.5	6729	5349
ByteTrack [33]	Train	62.8	69.8	51.2	94.1	77.1	85.6	3267	4499
OC-SORT [5]	Train	71.9	72.2	59.8	94.5	86.4	92.4	3093	3474
ByteTrack [33]	Train+Val	64.1	71.4	52.3	95.9	78.5	85.7	3089	4216
OC-SORT [5]	Train+Val	73.7	74.0	61.5	**96.5**	**88.5**	**92.7**	2728	3144
MixSort-Byte [7]	Train+Val	65.7	74.1	54.8	96.2	78.8	85.7	**2472**	4009
MixSort-OC [7]	Train+Val	74.1	74.4	62.0	**96.5**	**88.5**	92.7	2781	3199
Deep-EIoU (Ours)	Train	74.1	75.0	63.1	95.1	87.2	92.5	3066	3471
Deep-EIoU (Ours)	Train+Val	**77.2**	**79.8**	**67.7**	96.3	88.2	92.4	2659	**3081**

Table 2. The performance comparison between different state-of-the-art trackers on the SportsMOT test sets. Our algorithm outperforms all the other previous tracking algorithms and achieves SOTA performance in several major evaluation metrics including HOTA, IDF1, and AssA. The evaluation results besides BoT-SORT are taken from the number reported in the SportsMOT dataset paper [7]. While BoT-SORT is evaluated based on their official code [1].

Tracker	HOTA	AssA	DetA
DeepSORT [28]	69.552	58.668	82.628
ByteTrack [33]	71.500	60.718	84.342
BoT-SORT [1]	76.999	63.447	93.525
OC-SORT [5]	78.091	64.687	94.273
Deep-EIoU (Ours)	**85.443**	**73.567**	**99.236**

Table 3. Performance comparison of different tracking methods using oracle detections on the SoccerNet-Tracking [6] test set. The performance of DeepSORT and ByteTrack are reported from the SoccerNet-Tracking dataset paper [6]. While BoT-SORT and OC-SORT are evaluated using their official code.

Method	ReID	ISU	LI	HOTA (↑)
Baseline	-	-	-	71.403
-	✓			75.266
-	✓	✓		77.205
-	✓	✓	✓	77.220

Table 4. We evaluate the Deep-EIoU baseline with different settings on the SportsMOT test set. Including using the ReID model for association based on appearance, Iterative Scale-Up (ISU) process and using Linear Interpolation (LI) as post-processing for our method.

diction and detection will be expanded in the tracking process following the ExpansionIoU. The experiment results in Table 5 demonstrate that after directly replacing IoU with EIoU, these two classic Kalman filter-based trackers increase their performance by a large margin in HOTA, AssA, and DetA. This demonstrates that ExpansionIoU can also be applied as a plug-and-play trick for Kalman filter-based tracker to boost the tracking performance.

Tracker	w/ EIoU	HOTA	AssA	DetA
ByteTrack		62.8	51.2	77.1
ByteTrack	✓	67.5	54.4	83.9
BoT-SORT		68.7	55.9	84.4
BoT-SORT	✓	71.3	60.2	84.5

Table 5. We evaluate two classic Kalman filter-based tracking algorithms including ByteTrack [33] and BoTSORT [1] on the SportsMOT test set. Experiment results show that the Kalman filter-based tracker can also be benefited from incorporating ExpansionIoU during the tracking process.

4.10. Limitations

While our algorithm provides a robust and practical solution for online multi-object tracking in sports scenarios, it does have its limitations, including the absence of an offline post-processing trajectories refinement method. Such methods could involve a post-processing approach [14] or a strong memory buffer [26], which would be valuable in handling edge cases where sports players temporarily exit and re-enter the camera's field of view. It is worth noting that exploring and integrating offline refinement techniques in the future could potentially enhance the overall performance and extend the applicability of our approach beyond short-term tracking scenarios.

Another concern of Deep-EIoU is its relatively slower running speed when compared with motion-based trackers. Despite delivering significantly enhanced performance, the integration of the appearance-based tracking-by-detection framework, which involves a detector and a ReID model, introduces additional computational cost. The current Deep-EIoU pipeline achieves around 14.6 FPS on a single Nvidia RTX 4080 GPU, which is slower compared to motion-

Figure 6. Visualization results of Deep-EIoU from random sampled clips of SportsMOT dataset (row 1 to 3) and SoccerNet-Tracking dataset (row 4 to 5). With the iterative scale-up ExpansionIoU and deep features association, our algorithm can achieve robust multi-athlete tracking under severe occlusion conditions in multiple diverse sports scenarios including basketball, football, and volleyball. More visualization results can be found in supplementary material.

based method. It's worth noting that transitioning to a more lightweight detector and ReID model has the potential to significantly boost operational speed.

5. Conclusions

In this paper, we proposed Deep-EIoU, an iterative scale-up ExpansionIoU and deep features association method for multi-object tracking in sports scenarios, which achieves competitive performance on two large-scale multi-object sports player tracking datasets including SportsMOT and SoccerNet-Tracking. Our method successfully tackles the challenges of irregular movement during multi-object tracking in sports scenarios and outperforms the previous tracking algorithms by a large margin.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[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.

[9] Tobias Fischer, Jiangmiao Pang, Thomas E Huang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *arXiv preprint arXiv:2210.06984*, 2022.

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

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[12] Hsiang-Wei Huang, Cheng-Yen Yang, and Jenq-Neng Hwang. Multi-target multi-camera vehicle tracking using transformer-based camera link model and spatial-temporal information. *arXiv preprint arXiv:2301.07805*, 2023.

[13] Hsiang-Wei Huang, Cheng-Yen Yang, Zhongyu Jiang, Pyong-Kun Kim, Kyoungoh Lee, Kwangju Kim, Samartha Ramkumar, Chaitanya Mullapudi, In-Su Jang, Chung-I Huang, et al. Enhancing multi-camera people tracking with anchor-guided clustering and spatio-temporal consistency id re-assignment. *arXiv preprint arXiv:2304.09471*, 2023.

[14] Hsiang-Wei Huang, Cheng-Yen Yang, Samartha Ramkumar, Chung-I Huang, Jenq-Neng Hwang, Pyong-Kun Kim, Kyoungoh Lee, and Kwangju Kim. Observation centric and central distance recovery for athlete tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 454–460, 2023.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

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

[22] Robert Stone et al. Centertrack: An ip overlay network for tracking dos floods. In *USENIX Security Symposium*, volume 21, page 114, 2000.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Computer Vision–ECCV 2020: 16th European Conference,*

Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, pages 107–122. Springer, 2020.

[28] 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.

[29] Cheng-Yen Yang, Hsiang-Wei Huang, Zhongyu Jiang, Heng-Cheng Kuo, Jie Mei, Chung-I Huang, and Jenq-Neng Hwang. Sea you later: Metadata-guided long-term re-identification for uav-based multi-object tracking. *arXiv preprint arXiv:2311.03561*, 2023.

[30] Cheng-Yen Yang, Alan Yu Shyang Tan, Melanie J. Underwood, Charlotte Bodie, Zhongyu Jiang, Steve George, Karl Warr, Jenq-Neng Hwang, and Emma Jones. Multi-object tracking by iteratively associating detections with uniform appearance for trawl-based fishing bycatch monitoring, 2023.

[31] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4799–4808, 2023.

[32] 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.

[33] 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, 2021. arXiv preprint arXiv:2110.06864.

[34] 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.

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

[36] Zhonghan Zhao, Wenhao Chai, Shengyu Hao, Wenhao Hu, Guanhong Wang, Shidong Cao, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. A survey of deep learning in sports applications: Perception, comprehension, and decision. *arXiv preprint arXiv:2307.03353*, 2023.

[37] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015.

[38] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.

[39] 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.

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