D^3: Duplicate Detection Decontaminator for Multi-Athlete Tracking in Sports Videos

Rui He[1[0000−0002−5138−1654]], Zehua Fu[1,2[0000−0002−3639−4406]], Qingjie Liu[1,2⋆[0000−0002−5181−6451]], Yunhong Wang[1[0000−0001−8001−2703]], and Xunxun Chen[3[0000−0002−9481−4819]]

[1] Laboratory of Intelligent Recognition and Image Processing (IRIP Lab), Beihang University (BUAA), Xueyuan Road No.37, Haidian District, Beijing, China
[2] Hangzhou Innovation Institute, Beihang University, Hangzhou, China
[3] National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT or CNCERT/CC), Beijing, China
{heruihr,zehua_fu,qingjie.liu,yhwang}@buaa.edu.cn, cxx@cert.org.cn

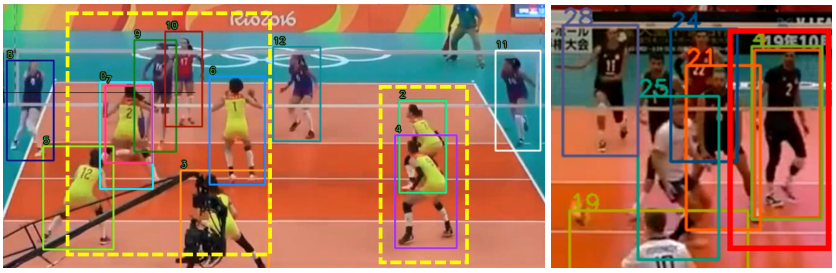
Abstract. Tracking multiple athletes in sports videos is a very challenging Multi-Object Tracking (MOT) task, since athletes often have the same appearance and are intimately covered with each other, making a common occlusion problem becomes an abhorrent duplicate detection. In this paper, the duplicate detection is newly and precisely defined as occlusion misreporting on the same athlete by multiple detection boxes in one frame. To address this problem, we meticulously design a novel transformer-based Duplicate Detection Decontaminator (D^3) for training, and a specific algorithm Rally-Hungarian (RH) for matching. Once duplicate detection occurs, D^3 immediately modifies the procedure by generating enhanced box losses. RH, triggered by the team sports substitution rules, is exceedingly suitable for sports videos. Moreover, to complement the tracking dataset that without shot changes, we release a new dataset based on sports video named RallyTrack. Extensive experiments on RallyTrack show that combining D^3 and RH can dramatically improve the tracking performance with 9.2 in MOTA and 4.5 in HOTA. Meanwhile, experiments on MOT-series and DanceTrack discover that D^3 can accelerate convergence during training, especially saving up to 80 percent of the original training time on MOT17. Finally, our model, which is trained only with volleyball videos, can be applied directly to basketball and soccer videos, which shows the priority of our method. Our dataset is available at https://github.com/heruihr/rallytrack.

Keywords: Multi-Athlete Tracking · Multi-Object Tracking · Transformer.

1 Introduction

Sports video analysis possesses wide application prospects and is currently receiving plenty of attention from academia and industry. Scene understanding

⋆ Corresponding Author



(a) (b)

Fig. 1. A labeled sample in RallyTrack and the duplicate detection problem occurred in a TransTrack model: (a) a labeled sample with heavy occlusions, label_7 is totally covered by label_0 in the left dash box; (b) red box shows duplicate detection, the same individual is detected by two queries with two IDs.

in sports video can be utilized for data statistics[13], highlight extraction[30], tactics analysis[21]. Multi-Athlete Tracking (MAT)[20] is a basic task in sports video-based scene understanding, which occupies a pivotal position.

Unlike general Multi-Object Tracking (MOT)[29,9], in MAT, different athletes generally share a high similarity in appearance and they often have a diversity of action changes and abrupt movements. The former difficulty, from our observation, turns a common occlusion problem in MOT to duplicate detection in sports video, which is defined in this paper as occlusion misreporting on the same object by multiple predictions in the same frame. The latter one leads to objects detetion missing, which often accompany with duplicate detection. In contrast to general person-based MOT, Figure 1(a) displays the difficulties with two yellow dash boxes. These two main difficulties make MAT a challenging task.

Then Figure 1(b) illustrates duplicate detection with a red box, which may be caused by two possibilities. One is that all athletes are detected, and an athlete being repeatedly detected is treated as occluding an additional invisible athlete. The other is that not all athletes are detected, some detections are missing while someone is repeatedly detected. To address this issue, we design a Duplicate Detection Decontaminator (D^3) , which can keep watch on the training procedure. Once a duplicate detection occurs, D^3 can generate additionally enhanced self-GIoU[35] losses during training. Then the losses will be gradually backpropagated to force the duplicate detecting boxes to keep away from each other. When duplicate detection disappears, D^3 would not produce loss anymore. We also offer a specific matching algorithm called Rally-Hungarian (RH) algorithm for MAT, which is triggered by the substitution rule in team sports like volleyball.

What is more, to make up for the lack of shot change, a new dataset namely RallyTrack is annotated, which is based on a scene of sports videos, and Figure 1(a) is a labeled sample. Unlike videos commonly used in the scientific research of MAT, live sports videos include shot changes as sports video streaming always uses multiple cameras. Although all athletes remain on the scene after each

shot change, the association becomes challenging as it is hard to predict the trajectories of athletes. However, there are a considerable amount of sports videos available online. Making use of those massive data can help improve athletes' competitiveness, e.g., using live sports videos for tactical analysis. Therefore, building a MAT dataset with shot changes is both significant academically and practically.

Intensive experimental results on RallyTrack demonstrate the efficiency of D^3 and RH. During our experiments, we discover that duplicate detection is not only a prominent problem in MAT but also an unnoticed barrier hidden in MOT, which makes a model converge slowly. Experimental results on MOT17[29] show D^3 can save up to 80 percent of original training time. More experiments on MOT16[22], MOT20[9], and DanceTrack[37] also confirm the priority of D^3.

The main contributions of this study are as follows. (i) We design a Duplicate Detection Decontaminator (D^3) which supervises the training procedure to optimize detection and tracking boxes. (ii) We design a matching algorithm called Rally-Hungarian (RH) for MAT to further improve tracking result. (iii) We annotate a new dataset named RallyTrack, which is based on scenes of sports videos, to make up for the lack of videos without shot change. (iv) We perform extensive experiments to demonstrate and verify that the proposed method improves the tracking performance on MAT with a total enhancement of 9.2 for MOTA and 4.5 for HOTA, and D^3 can accelerate training convergence on MOT.

2 Related Work

2.1 Multiple Object Tracking Datasets

Human-based Datasets. Concentrating on variant scenarios, a large number of multiple object tracking datasets have been collected, and human tracking datasets accounted for a big proportion. PETS [11], MOT15 [22], MOT17 [29] and MOT20 [9] datasets become popular in this community. MOT datasets mainly contain a handful of pedestrian videos, which are limited to regular movements of objects and distinguishable appearances. As a consequence of that, multiple object tracking could be easily achieved with the association by pure appearance matching [31]. More recently DanceTrack [37] is proposed as being expected to make research rely less on visual discrimination and depend more on motion analysis. However, the background in DanceTrack is usually identical to the foreground so detecting is easy and tracking is hard. Collected from real and noisy sports videos, our dataset is both challenging in detecting and tracking.

Diverse Datasets. Besides, WILDTRACK [6] , Youtube-VOS [45], and MOTS [41] are proposed for diverse objectives. With the development of autonomous driving, KITTI [12] is interested in vehicles and pedestrians. Then larger scale autonomous driving datasets BDD100K[48] and Waymo[36] are published. Limiting by lanes and traffic rules, the motion patterns of objects in these datasets are even more regular than moving people. What is more, some datasets broaden their horizon on more diverse object categories, such as ImageNet-Vid [10] and TAO [8].

2.2 Object Detection in MOT

Tracking by detection. Object detection [24,33,34] develops so vigorously that a lot of methods would like to utilize powerful detectors to pursue higher tracking performance. RetinaTrack [26] and ChainedTracker [32] apply the one-stage object detector RetinaNet [24] for tracking. For its simplicity and efficiency, CenterNet [54] becomes a popular detector adopted by CenterTrack [53] and FairMOT [51] . The YOLO series detectors [33] are also put to use by TransMOT [7] due to their excellent balance of accuracy and speed. Single image tracking is easy for these methods. However, as is pointed out by [39], when occlusion happens, missing and low-scoring detections would influence the quality of object linking. Therefore, the information of the previous frame is usually leveraged to enhance the video detection performance. Recently, Versatile Affinity Network (VAN) [23] is proposed to handle incomplete detection issues, affinity computation between target and candidates, and decision of tracking termination. [16] presents an approach injecting spatiotemporally derived information into convolutional AutoEncoder in order to produce a suitable data embedding space for multiple object tracking.

Joint-detection-and-tracking. Achieving detection and tracking simultaneously in a single stage is the destination of the joint-detection-and-tracking pipeline. Some early methods [7] utilize single object tracking (SOT) [3] or Kalman filter [19]to predict the location of the tracklets in the following frame and fuse the predicted boxes with the detection boxes. Then by combining the detection boxes and tracks , Integrated-Detection [52] boosts the detection performance. Recently, Tracktor [1] directly regress the previous frame tracking boxes to provide tracking boxes on the current step. From a shared backbone, JDE [42] and FairMOT [51] learn the object detection task and appearance embedding task in the meantime. Different from CenterTrack [53] localizing objects by tracking-conditioned detection and predicting their offsets to the previous frame, ChainedTracker [32] chains paired bounding boxes estimated from overlapping nodes, in which each node covers two adjacent frames. More recently, transformer-based [40] detectors like DETR [5,55] are adopted by several methods, such as TransTrack [38],TrackFormer [28] , and MOTR [49] . Our method also follows this structure to utilize the similarity with tracklets to strengthen the reliability of detection boxes.

2.3 Data Association

Tracking by matching appearance. Appearance similarity is useful in long-range matching and serves as a linchpin in many multi-object tracking methods. DeepSORT [43] adopts a stand-alone Re-ID model to extract appearance features from the detection boxes. POI [47] achieves excellent tracking performance depending on the high-quality detection. Recently, because of their simplicity and efficiency, joint detection and Re-ID models, such as RetinaTrack [26], QuasiDense (QDTrack) [31], JDE [42], FairMOT [51], become more and more prevalent.

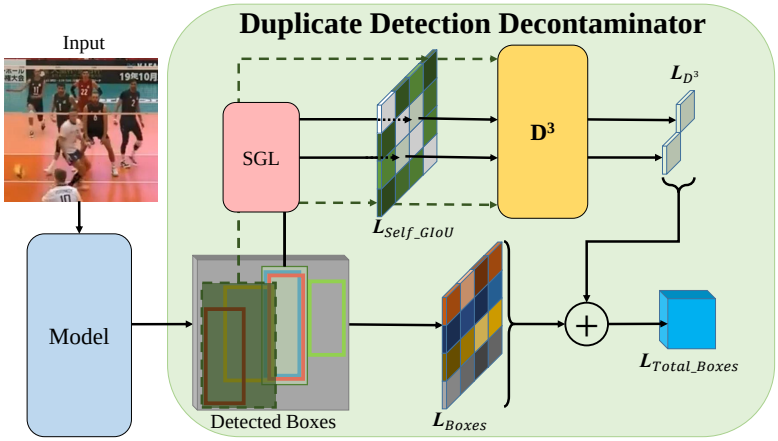


Fig. 2. In the training stage, a matrix L_{Self_GIoU} is constructed according to the detected boxes of the input frame t by Self-GIoU Loss (SGL) function. Then D^3 will set a Lower Bound (LB) to check L_{Self_GIoU}. Values lower than LB in the matrix will be regarded as duplicate detection. The values are output and added to the boxes loss L_{Boxes} as a total boxes loss L_{Total_Boxes} to be backpropagated. If there is no duplicate detection, D^3 will do nothing.

Tracking with motion analysis. Tracking objects by estimating their motion is a natural and intuitive idea. SORT [4] first adopts Kalman filter [19] to predict the location of the tracklets in the new frame, and then by Hungarian algorithm [46] computes the IoU between the detection boxes and the predicted boxes as the similarity for tracking. STRN [44] presents a similarity learning framework between tracks and objects. Tracking by associating almost every detection box instead of only the high score ones, for the low score detection boxes, ByteTrack [50] utilizes their similarities with tracklets to recover true objects and filter out the background detections. Recently attention mechanism [40] can directly propagate boxes between frames and perform association implicitly. TransTrack [38] is designed to learn object motions and achieves robust results in cases of large camera motion or low frame rate.

3 Duplicate Detection Decontaminator and Rally-Hungarian Algorithm

In this section, the working mechanism of Duplicate Detection Decontaminator (D^3) and Rally-Hungarian (RH) matching algorithm will be introduced in detail respectively. In a transformer-based joint-detection-and-tracking model, objects in an image are detected by harnessing learned object queries, which is a set of learnable parameters trained together with all other parameters in the network. While training a model, duplicate detection appears and D^3 will unroll its power

then. Then RH, a box IoU matching method, is utilized to obtain the final tracking result by associating object queries and tracking queries.

3.1 Duplicate Detection Decontaminator

At training stage as shown in Figure 2, denote $B = \{\mathbf{b}_i | i = 1, \cdots, N\}$ as the boxes set of individuals in the middle output of input frame t, where $\mathbf{b}_i = (x_i^1, y_i^1, x_i^2, y_i^2)$ indicates the top-left corner (x_i^1, y_i^1) and bottom-right corner (x_i^2, y_i^2) of ith individual. Then applying the concept of Generalized Intersection over Union (GIoU) [35], we get Self-GIoU from B, where the element in $GIoU(B, B)$ is formulated as follows:

$$GIoU(\mathbf{b}_i, \mathbf{b}_j) = \frac{|\mathbf{b}_i \cap \mathbf{b}_j|}{|\mathbf{b}_i \cup \mathbf{b}_j|} - \frac{|V \setminus (\mathbf{b}_i \cup \mathbf{b}_j)|}{|V|} = IoU(\mathbf{b}_i, \mathbf{b}_j) - \frac{|V \setminus (\mathbf{b}_i \cup \mathbf{b}_j)|}{|V|} \tag{1}$$

where V is the smallest convex hull that encloses both $\mathbf{b}_i$ and $\mathbf{b}_j$, IoU means Intersection over Union. Then a matrix L_{Self_GIoU} is constructed by Self-GIoU Loss (SGL) function as follows:

$$L_{Self\_GIoU} = SGL(B) = 1 - GIoU(B, B) \tag{2}$$

L_{Self_GIoU} is a symmetric matrix in which the elements on the diagonal of the matrix are all 0, as painted white in Figure 2. Then D^3 will set a Lower Bound (LB) to check L_{Self_GIoU}. Once a value of the element in L_{Self_GIoU} is lower than the LB, which means duplicate detection happens, D^3 will output the value and add it to the detected boxes loss as a total boxes loss to be backpropagated. If there is no duplicate detection, D^3 will do nothing. The mechanism of D^3 is as follows:

$$L_{D^3} = D^3(L_{Self\_GIoU}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} l_{ij}, \quad l_{ij} < LB \tag{3}$$

$$L_{Total\_Boxes} = L_{Boxes} + L_{D^3} \tag{4}$$

where l_{ij} is an element located at ith row and jth column in $L_{Self-GIoU}$. l_{ij} is equal to l_{ji} in a symmetric matrix so the output of D^3 should be divided by two. When the model is equipped D^3, duplicate detection may be within limits. However, in sports video, the quality of MOT could go a step further while making use of some special rules of sports, which pedestrian video is not in the possession of.

3.2 Rally-Hungarian Algorithm

Although the Hungarian algorithm can still work, its limitation is shown when applied to sports videos. So Rally-Hungarian (RH) algorithm is provided and its overview is shown in Figure 3. RH models the substitution rule of team sports.

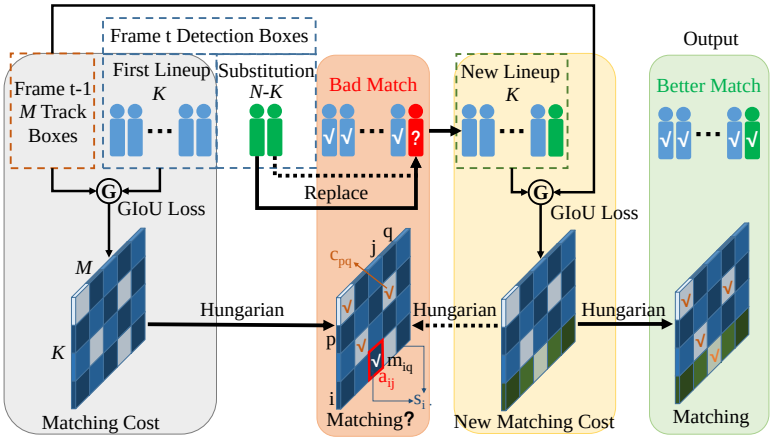


Fig. 3. Through a tracking model, detection boxes of frame t and track boxes of frame $t-1$ are acquired. Then detection boxes are split into Lineup and Substitution. A rally of constructing matching cost matrix by computing GIoU loss from Lineup and track boxes, getting matching pairs by applying Hungarian algorithm on the matrix, replacing the bad matching item with Substitution to form new Lineup is executed, which is called Rally-Hungarian (RH).

The players who are on the court are called Lineup, and the others from the same team are called Substitution. Usually, if a Lineup player performs not good, he or she will be replaced by a Substitution player. Then RH is introduced in detail.

Through a tracking model, detection boxes of frame t which are denoted as $B_{det} = \{\mathbf{b}_n | n = 1, \cdots, N\}$, in which the elements are sorted in descending order by detection score, and track boxes of frame $t-1$ as $B_{track} = \{\mathbf{b}_j | j = 1, \cdots, M\}$ are acquired, as the definition of boxes set applied in Section 3.1. According to the substitution rule, we split detection boxes set B_{det} to B_{lineup} and B_{sub}. The top K elements in B_{det} are regarded as Lineup $B_{lineup} = \{\mathbf{b}_i | i = 1, \cdots, K\}$ and the rest elements as Substitution $B_{sub} = \{\mathbf{b}_k | k = K+1, \cdots, N\}$. We provide a mathematical explanation of setting limitation K in RH, please refer to supplementary 1. Then we construct matching cost matrix $\mathbf{C} \in \mathbb{R}^{K \times M}$ by computing GIoU loss from B_{lineup} and B_{track} as follows:

$$\mathbf{C} = 1 - GIoU(B_{lineup}, B_{track}) \tag{5}$$

where GIoU is the same as equation (1). Then utilizing Hungarian Algorithm on $\mathbf{C}$, we could acquire a set of matching indices pairs $P = \{(i, j) | i \in [1, K]; j \in [1, M]\}$ as follows:

$$P = Hungarian(\mathbf{C}) \tag{6}$$

P are labeled by check marks in Figure 3. If $\mathbf{b}_i$ and $\mathbf{b}_j$ belong to one individual, item c_{ij} in $\mathbf{C}$ should be a relatively small value as the light color square shows,

which means an individual is tracked. However, if an abnormality is chosen, as marked in red, a row where the abnormality is occupying should be replaced.

Here, we explain why rows with abnormalities can be replaced. In Figure 3, according to the Hungarian algorithm, a abnormality a_{ij} in $\mathbf{C}$ at row i and column j is the best match. It means that in $\mathbf{C}$, of all the mismatched columns $s_{i\cdot}$ in row i, the value is the minimum, which can be written as:

$$a_{ij} = min\{s_{i\cdot}\} \tag{7}$$

It is also not considered a new target in sports video. Two inferences can then be drawn. First, some duplicate detections are not eliminated by D^3. Then there must be a value m_{iq} smaller than the abnormality that exists among all the matching columns in the row i,

$$m_{iq} < a_{ij} \tag{8}$$

and the smaller value does not match. That is to say, in column q, two similar values exist in row i and another row p,

$$m_{iq} \approx c_{pq} \tag{9}$$

which means row i is duplicate detection and can be replaced. Secondly, this abnormality is exactly the minimum value of this row, indicating a bad quality of the matching. As a consequence, a row, or a detection box, with an abnormality could be replaced.

Then the bad matching detection box in B_{lineup}, regarded as B_{bad}, could be replaced by a substitution in B_{sub}, and a new lineup set B_{new} is composed as follows:

$$B_{new} = (B_{lineup} \setminus B_{bad}) \cup B_{sub} = (B_{lineup} \setminus \mathbf{b}_i) \cup \mathbf{b}_k \tag{10}$$

Looping equations (5), (6), (10) as the dash arrows until each $\mathbf{C}_{ij}$ becomes acceptable or the B_{sub} is empty. In the end, we get a better match pair set. In the field of volleyball, a rally means a round will not stop until the ball touch floor, like a loop. So we name our matching strategy as Rally-Hungarian (RH) algorithm, and "R" may have a dual meaning of "Replace".

4 RallyTrack Dataset

There are plenty of MOT Datasets as we have mentioned in Section 2.1. However, there are few datasets for MAT. Driven by this observation, a question arises: Is anything difficult while exploring MAT? To discover the mystery in MAT, we annotate a RallyTrack dataset based primarily on sports videos for the MAT task as shown in Figure 4. In this section, RallyTrack Dataset will be introduced in detail and our labeling method is provided in our supplementary material 2.

In RallyTrack, videos are from different views, broadcast or fixed, and different gender, men or women, of volleyball games. To guarantee training data and

(a) (b) (c)

Fig. 4. Some samples in RallyTrack: (a) broadcast view of men's game; (b) fixed view of men's game; (c) broadcast view of women's game.

Table 1. Datasets comparison between MOT17 and RallyTrack. F/V means frames per video. O/F means objects per frame. T/V means tracks per video.

Dataset	Subset	Videos	Frames	F/V	Objects	O/F	Tracks	T/V
MOT17	Train	7	5316	759.4	85828	16.1	546	78
	Test	7	5919	845.6	-	-	-	-
	Total	14	11235	802.5	-	-	-	-
RallyTrack	Train	10	8104	810.4	68449	8.5	122	12.2
	Test	10	9757	975.7	91126	9.3	126	12.6
	Total	20	17861	893.1	159575	8.9	248	12.4

test data are not crossed, only will games from different Series be set as train and test. For example, if both games come from Rio 2016 Olympic Games, they should be put into a train set or test set together, even if each team is different. Some details of RallyTrack are then displayed in Table 1. All of our data are labeled in MOT17 annotation format. As the test set's ground truths of MOT17 are not published, only the train set is calculated. In this table, column F/V refers to the number of frames showing more frames in RallyTrack than in each MOT17 video. Column O/F means objects per frame which show that individuals in RallyTrack are less than MOT17. Column T/V means tracks per video which show that trajectories in RallyTrack are also less than MOT17. However, given the overall situation of O/F and T/V, O/F is closer to T/V in RallyTrack than in MOT17, suggesting that RallyTrack has a longer personal trajectory than MOT17.

5 Experimental Results

5.1 Experimental Setup

We evaluate D^3 on benchmarks: RallyTrack, MOT17, MOT16, MOT20, and DanceTrack. Following previous practice [53,38], we split all the training sets of the MOT-series into two parts, one for training and the other for validation. The operation is samely applied on RallyTrack where half of the train set will be used and the whole test set will be tested. The widely-used MOT metrics set

Table 2. Experiments on RallyTrack. Our method makes an amazing 9.2 promotion on MOTA, 7.0 on IDF1 and 4.5 on HOTA to baseline TransTrack (TT). Best in bold.

Model	MOTA↑	IDF1↑	MOTP↑	MT↑	FP↓	FN↓	IDS↓	HOTA↑	DetA↑	AssA↑
TT [38]	59.5	28.8	77.8	70.6	15370	19489	2049	27.9	51.9	15.2
TT+RH	62.0	33.3	77.8	66.7	12557	20310	**1788**	30.2	52.3	17.7
TT+D^3	66.4	29.7	78.1	**78.6**	13676	**14848**	2107	29.2	55.8	15.5
TT+D^3+RH	**68.7**	**35.8**	**78.1**	77.0	**11359**	15350	1847	**32.4**	**56.3**	**18.9**

[2] is adopted for quantitative evaluation where multiple objects tracking accuracies (MOTA) is the primary metric to measure the overall performance. What is more, the higher order tracking accuracy (HOTA) [27,18], which explicitly balances the effect of performing accurate detection, association and localization into a single unified metric for comparing trackers, is also applied. While evaluating RH, only RallyTrack is used.

For a fair comparison, we maintain most of the settings in TransTrack [38], such as ResNet-50 [15] network backbone, Deformable DETR [55] based transformer structure, AdamW [25] optimizer, batch size 16. The initial learning rate is 2e-4 for the transformer and 2e-5 for the backbone. The weight decay is 1e-4. All transformer weights are initialized with Xavier-init [14]. The backbone model is pre-trained on ImageNet [10] with frozen batch-norm layers [17]. Data augmentation includes random horizontal, random crop, scale augmentation, and resizing the input images whose shorter side is by 480-800 pixels while the longer side is by at most 1333 pixels. When the model is trained for 150 epochs, the learning rate drops by a factor of 10 at the 100th epoch.

5.2 Experiments on RallyTrack and Others

Our models are evaluated on RallyTrack as shown in Table 2. In this table, the original TransTrack (TT) [38] model based on Deformable Transformer [55] is regarded as a baseline, Rally-Hungarian (RH) algorithm and Duplicate Detection Decontaminator (D^3) could be evaluated respectively or jointly. In this table, TT with both D^3 and RH gets a stunning 9.2 rating on MOTA, 7.0 on IDF1 and 4.5 on HOTA to baseline. The results show that our methods are not only good at detecting multiple athletes but also associating them. It is mainly caused by decontaminating the duplicate detections and many athletes are correctly detected and tracked.

MOT17 is another dataset mainly used to measure the effectiveness of D^3 as shown in Table 3. In this dataset, the main function of our approach is to reduce training time. The hyperparameters in the first column mean Lower Bound (LB) in D^3. LB is chosen according to different self-GIoU losses from different datasets. Different self-GIoU losses are caused by different resolutions of videos. In this table, by actively eliminating duplicate detection, D^3 can save 80 percent of TT's

Table 3. Experiments on MOT17, MOT16, MOT20, and DanceTrack. Our method converges faster. Best in bold.

D^3	Epoch	MOTA↑	IDF1↑	MOTP↑	MT↑	FP↓	FN↓	IDS↓	HOTA↑	DetA↑	AssA↑
w/o	150	65.1	63.6	81.9	36.8	1918	16440	**438**	52.6	54.0	51.7
0.010	150	65.3	62.9	82.2	38.3	1849	**16358**	457	53.0	54.4	52.1
w/o	30	64.9	62.6	82.0	36.3	1862	16537	477	52.1	53.9	50.7
0.010	**30**	**65.3**	**63.6**	**82.2**	**38.6**	**1833**	16398	480	**53.4**	**54.5**	**52.8**
mot16											
w/o	30	64.1	61.3	81.8	40.7	**2434**	16186	**544**	50.9	53.6	48.9
0.011	**30**	**65.3**	**61.6**	81.8	40.7	2578	**15328**	601	**52.3**	**55.0**	**50.3**
mot20											
w/o	30	72.5	63.2	82.9	51.6	12882	153K	2978	52.4	59.4	46.3
0.017	**30**	**73.2**	**64.6**	82.9	**53.4**	**12831**	**149K**	**2808**	**53.6**	**60.1**	**47.9**
DT											
w/o	50	76.5	39.4	85.2	**70.7**	19087	29130	4795	**38.9**	**66.8**	22.9
0.012	50	76.3	37.4	84.8	68.5	19432	**28947**	5026	37.1	66.4	21.0
w/o	25	**76.6**	37.6	**85.2**	70.0	18710	29348	**4685**	38.1	67.1	21.8
0.012	**25**	76.5	**39.4**	84.9	67.8	**18518**	29557	4808	38.7	66.2	**22.9**

training time, making the model converge faster from 150 down to 30 epochs. Instead, too many training epochs can lead to overfitting. We then demonstrated the priority of our approach by experimenting directly with MOT16 and MOT20 in the same setting as MOT17 for only 30 epochs. DanceTrack (DT) dataset is also measured with training on train set and testing on val set. As shown in Table 3 only trained in 50 epochs could our method save 50 percent of the original training time and almost maintain the basic performance.

Datasets and solutions are massive for MOT after long-term development while it is not for MAT. We hope to provide a paradigm for MOT methods to easily extend to MAT. So D^3 is proposed as a connection between them. D^3 retains an almost complete structure of TT, allowing TT to expand for MAT (Table 2) while maintaining the original MOT capabilities (Table 3).

5.3 Ablation Study

In this section, we conduct a comprehensive ablation study for the proposed D^3 and RH.

What is the Shortest Training Time and the Best Lower Bound of D^3? Training epochs and lower bound (LB) are two key factors for D^3 networks. During training, short training time leads to non-convergence and long training time leads to overfitting. Then a small LB has little effect and a large LB leads to the elimination of non-duplicate detection. We verify the impact in training with

Table 4. Training Time and Lower Bound. Best in bold.

Model	LB	Epoch	MOTA↑	IDF1↑	MT↑	FP↓	FN↓	IDS↓	HOTA↑	DetA↑	AssA↑
TT	-	150	59.4	28.7	69.8	15520	19426	2052	28.1	51.9	15.4
TT+D^3	0.010	150	61.9	29.2	73.0	15393	17166	2164	28.9	53.4	15.9
TT+D^3	0.011	150	66.3	29.5	78.6	13806	**14750**	2109	29.2	55.8	15.5
TT+D^3	0.012	150	61.8	29.0	74.6	15954	16540	2309	28.8	53.0	16.0
TT	-	40	59.5	28.8	70.6	15370	19489	**2049**	27.9	51.9	15.2
TT+D^3	0.011	40	**66.4**	**29.7**	**78.6**	**13676**	14848	2107	**29.2**	**55.8**	**15.5**

Table 5. Age L, Top K and Replacement of RH.

	L	K	Replace	MOTA	IDF1	HOTA
TT+D^3	32	-	No	66.4	29.7	29.2
TT+D^3	80	-	No	66.3	32.2	30.3
TT+D^3+RH	32	12	No	67.7	30.9	30.3
TT+D^3+RH	80	12	No	67.7	33.2	31.0
TT+D^3+RH	80	12	Yes	**68.7**	**35.8**	**32.4**

different training epochs and LB settings. Table 4 shows that using 40 training epochs and 0.011 LB brings the best performance in terms of MOTA, IDF1, and HOTA. On one hand, D^3 makes more correct predictions shown by FP and FN with a desirable gap. On the other hand, LB in D^3 should be carefully set to determine whether duplicate detection exists.

What is the Best Age L of RH? Should RH Set Top K and Conduct Replacement? First, age L means that if a tracking box is unmatched, it keeps as an "inactive" tracking box until it remains unmatched for L consecutive frames. Inactive tracking boxes can be matched to detection boxes and regain their ID. Following [38], we choose L=32 and then lengthen L to 80. Because in sports videos individuals who are always on the court will reappear in an image. Second, we set top K=12 as the data are based on volleyball videos. Finally, whether replace with substitution is also evaluated. Tabel 5 shows that using age L=80, setting a limitation K=12, and conducting replacement bring the best performance.

What is the Best Replacing Strategy of RH? Different replacing strategies (RS) may lead to different tracking performances. So 5 different RS of the RH algorithm are examined as shown in Table 6. In this table, we assume that there are p to-be replaced items in B_{bad} and q items in Substitution. Delete No. means the number of removed items in B_{bad}, and "1st Bad" means deleting the first item in B_{bad}. Replace No. means the number of being replaced items in B_{sub}. As the elements in B_{sub} are already sorted in descending order by detection score, the first one has the highest detection score in B_{sub}, which is marked

Table 6. Replacing Strategies of RH.

	Delete No.	Replace No.	Complexity	MOTA	IDF1	HOTA	FPS
RS1	p	1 (1st Score)	$O(p)$	67.5	31.9	30.4	6.41
RS2	p	1 (1st Good)	$O(pq)$	67.5	31.8	30.2	6.43
RS3	p	$min\{p,q\}$	$O(p \cdot min\{p,q\})$	67.6	32.2	30.4	6.33
RS4	1 (1st Bad)	1 (1st Score)	$O(1)$	68.4	35.5	**32.5**	**6.52**
RS5	1 (1st Bad)	1 (1st Good)	$O(q)$	**68.7**	**35.8**	32.4	6.44

Table 7. Experiments on basketball and soccer.

Basketball	MOTA↑	IDF1↑	MOTP↑	MT↑	FP↓	FN↓	IDS↓	HOTA↑	DetA↑	AssA↑
TT [38]	39.7	12.5	72.9	20.0	3995	3396	636	14.8	43.1	5.2
TT+RH	45.1	14.4	73.0	20.0	3079	3680	593	15.9	44.6	5.7
TT+D^3	53.2	12.5	74.7	20.0	2450	**3205**	605	15.2	**48.6**	4.8
TT+D^3+RH	**54.4**	**16.8**	**74.7**	20.0	**2166**	3420	**520**	**17.4**	48.0	**6.3**
Soccer										
TT [38]	**60.1**	21.2	79.3	33.3	1232	**4205**	**327**	23.2	**51.4**	10.5
TT+RH	59.2	**24.0**	**79.4**	**42.9**	1287	4260	354	**24.1**	50.9	**11.4**
TT+D^3	55.7	19.3	78.9	33.3	**1156**	4868	381	20.8	48.2	9.0
TT+D^3+RH	57.9	21.4	78.8	28.6	1175	4907	395	22.0	47.9	10.2

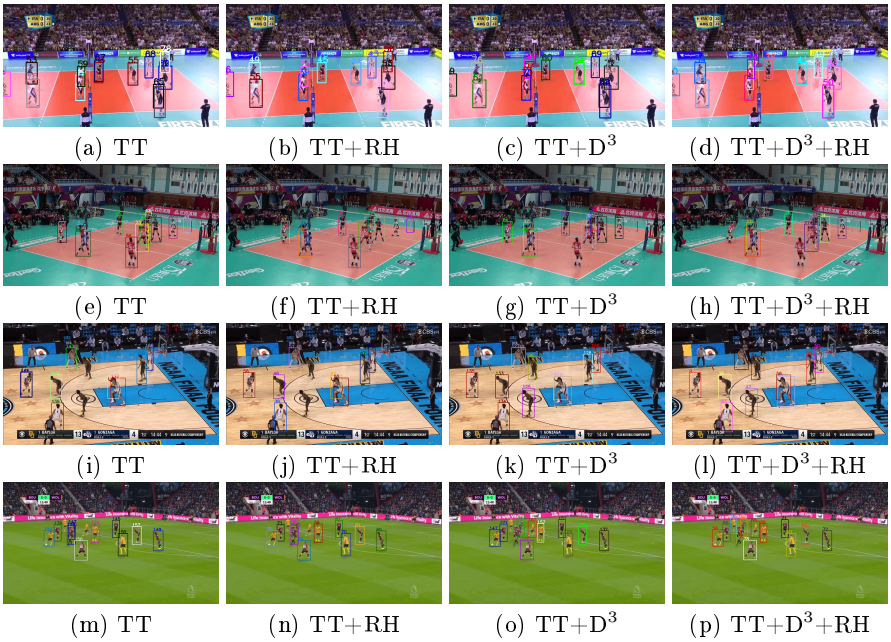
as "1st Score". When a bad item is replaced by a high score substitution, it is also able to get a bad match. So all the items in B_{sub} could be replaced to find a good match, and the first item composing a good match is marked as "1st Good". Then time complexity of each strategy is also analyzed. In this table, the total q is set 3, so the FPSes are close.

5.4 Details of basketball and soccer videos

Additionally, extending the RallyTrack dataset to other sports, we labeled 1484 frames of basketball and 1422 frames of soccer and tested them as shown in Table 7. 13384 objects are in basketball and 14802 objects in soccer. Results indicate that our method can be directly applied to basketball videos rather than soccer videos. The scene in basketball is more similar to volleyball than that in soccer, and occlusion is not serious in soccer mainly because the background is easily distinguishable and larger, as visualized in Figure 5.

5.5 Visualization

We visualize two examples tracked by four different tracking models as shown in Figure 5. In 5(a), heavy duplicate detections happen and an object is missing while using a base model TT; then in 5(b), with RH, some duplicate detections



(a) TT (b) TT+RH (c) TT+D^3 (d) TT+D^3+RH

(e) TT (f) TT+RH (g) TT+D^3 (h) TT+D^3+RH

(i) TT (j) TT+RH (k) TT+D^3 (l) TT+D^3+RH

(m) TT (n) TT+RH (o) TT+D^3 (p) TT+D^3+RH

Fig. 5. Visualization of frame No.942 in test_0100017 and frame No.715 in validation_0160025 tracked by four different tracking models, and the models are directly applied to basketball and soccer videos.

disappear; moreover in 5(c), when equipped D^3, the missing object is found; finally in 5(d), combining D^3 and RH could get the best and the clearest tracking result. Then the best model is directly applied to basketball by setting $N = 15$, $K = 10$, $q = 5$ in RH, and soccer by $N = 20$, $K = 15$, $q = 5$ for the court of soccer is so large that usually not all individuals are visible.

6 Conclusion

In this paper, to address duplicate detection in MAT, we design a Duplicate Detection Decontaminator (D^3) which supervises the training procedure.Then we design a Rally-Hungarian (RH) matching algorithm to go a step further on MAT. Experiments on our labeled RallyTrack show the priority of our methods. D^3 could also be utilized for saving training time on MOT17, MOT16, MOT20, and DanceTrack. Moreover, our model trained with volleyball data can be directly applied on other team sports videos like basketball or soccer, which may encourage more research exploring MAT applications.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant U20B2069 and the Fundamental Research Funds for the Central Universities.

References

1. Bergmann, P., Meinhardt, T., Leal-Taixé, L.: Tracking without bells and whistles. In: International Conference on Computer Vision. pp. 941–951 (2019)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. EURASIP J. Image Video Process. **2008** (2008)
3. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional siamese networks for object tracking. In: European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 9914, pp. 850–865 (2016)
4. Bewley, A., Ge, Z., Ott, L., Ramos, F.T., Upcroft, B.: Simple online and realtime tracking. In: International Conference on Image Processing. pp. 3464–3468 (2016)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (2020)
6. Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T.M., Lettry, L., Fua, P., Gool, L.V., Fleuret, F.: WILDTRACK: A multi-camera HD dataset for dense unscripted pedestrian detection. In: Computer Vision and Pattern Recognition. pp. 5030–5039 (2018)
7. Chu, P., Wang, J., You, Q., Ling, H., Liu, Z.: Transmot: Spatial-temporal graph transformer for multiple object tracking. arxiv **abs/2104.00194** (2021)
8. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: TAO: A large-scale benchmark for tracking any object. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 12350, pp. 436–454 (2020)
9. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I.D., Roth, S., Schindler, K., Leal-Taixé, L.: MOT20: A benchmark for multi object tracking in crowded scenes. arxiv **abs/2003.09003** (2020)
10. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition. pp. 248–255 (2009)
11. Ellis, A., Ferryman, J.M.: PETS2010 and PETS2009 evaluation of results using individual ground truthed single views. In: International Conference on Advanced Video and Signal-based Surveillance. pp. 135–142 (2010)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Computer Vision and Pattern Recognition. pp. 3354–3361 (2012)
13. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: A scalable dataset for action spotting in soccer videos. In: Computer Vision and Pattern Recognition Workshops. pp. 1711–1721 (2018)
14. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics. JMLR Proceedings, vol. 9, pp. 249–256 (2010)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition. pp. 770–778 (2016)
16. Ho, K., Kardoost, A., Pfreundt, F., Keuper, J., Keuper, M.: A two-stage minimum cost multicut approach to self-supervised multiple person tracking. In: Asian Conference on Computer Vision. Lecture Notes in Computer Science, vol. 12623, pp. 539–557. Springer (2020)

17. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. JMLR Workshop and Conference Proceedings, vol. 37, pp. 448–456 (2015)
18. Jonathon Luiten, A.H.: Trackeval. `https://github.com/JonathonLuiten/TrackEval` (2020)
19. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering **82D**, 35–45 (1960)
20. Kong, L., Huang, D., Wang, Y.: Long-term action dependence-based hierarchical deep association for multi-athlete tracking in sports videos. IEEE Trans. Image Process. **29**, 7957–7969 (2020)
21. Kong, L., Zhu, M., Ran, N., Liu, Q., He, R.: Online multiple athlete tracking with pose-based long-term temporal dependencies. Sensors **21**(1), 197 (2021)
22. Leal-Taixé, L., Milan, A., Reid, I.D., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arxiv **abs/1504.01942** (2015)
23. Lee, H., Kim, I., Kim, D.: VAN: versatile affinity network for end-to-end online multi-object tracking. In: Asian Conference on Computer Vision. Lecture Notes in Computer Science, vol. 12623, pp. 576–593. Springer (2020)
24. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: International Conference on Computer Vision. pp. 2999–3007 (2017)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
26. Lu, Z., Rathod, V., Votel, R., Huang, J.: Retinatrack: Online single stage joint detection and tracking. In: Computer Vision and Pattern Recognition. pp. 14656–14666 (2020)
27. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International Journal of Computer Vision pp. 1–31 (2020)
28. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
29. Milan, A., Leal-Taixé, L., Reid, I.D., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arxiv **abs/1603.00831** (2016)
30. Niu, Z., Gao, X., Tian, Q.: Tactic analysis based on real-world ball trajectory in soccer video. Pattern Recognit. **45**(5), 1937–1947 (2012)
31. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Computer Vision and Pattern Recognition. pp. 164–173 (2021)
32. Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 12349, pp. 145–161 (2020)
33. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arxiv **abs/1804.02767** (2018)
34. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Conference and Workshop on Neural Information Processing Systems. pp. 91–99 (2015)
35. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I.D., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Computer Vision and Pattern Recognition. pp. 658–666 (2019)

36. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Computer Vision and Pattern Recognition. pp. 2443–2451 (2020)
37. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
38. Sun, P., Jiang, Y., Zhang, R., Xie, E., Cao, J., Hu, X., Kong, T., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple-object tracking with transformer. arxiv **abs/2012.15460** (2020)
39. Tang, P., Wang, C., Wang, X., Liu, W., Zeng, W., Wang, J.: Object detection in videos by high quality object linking. IEEE Trans. Pattern Anal. Mach. Intell. **42**(5), 1272–1278 (2020)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Conference and Workshop on Neural Information Processing Systems. pp. 5998–6008 (2017)
41. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: MOTS: multi-object tracking and segmentation. In: Computer Vision and Pattern Recognition. pp. 7942–7951 (2019)
42. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 12356, pp. 107–122 (2020)
43. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: International Conference on Image Processing. pp. 3645–3649 (2017)
44. Xu, J., Cao, Y., Zhang, Z., Hu, H.: Spatial-temporal relation networks for multi-object tracking. In: International Conference on Computer Vision. pp. 3987–3997 (2019)
45. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B.L., Cohen, S., Huang, T.S.: Youtube-vos: Sequence-to-sequence video object segmentation. In: Computer Vision European Conference. Lecture Notes in Computer Science, vol. 11209, pp. 603–619. Springer (2018)
46. Yaw, H.: The hungarian method for the assignment problem. In: Naval Res Logist Quart (1955)
47. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: POI: multiple object tracking with high performance detection and appearance feature. In: European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 9914, pp. 36–42 (2016)
48. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: Computer Vision and Pattern Recognition. pp. 2633–2642 (2020)
49. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: European Conference on Computer Vision (2022)
50. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box (2022)
51. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vis. **129**(11), 3069–3087 (2021)

52. Zhang, Z., Cheng, D., Zhu, X., Lin, S., Dai, J.: Integrated object detection and tracking with tracklet-conditioned detection. arxiv **abs/1811.11167** (2018)
53. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 12349, pp. 474–490 (2020)
54. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arxiv **abs/1904.07850** (2019)
55. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)